Multimedia Event Detection Using Hidden Conditional Random Fields

Kimiaki Shirahama and Marcin Grzegorzek
Research Group for Pattern Recognition
University of Siegen, Germany
Hoelderlinstr. 3, D-57076 Siegen
{kimiaki.shirahama,marcin.grzegorzek}@uni-siegen.de

Kuniaki Uehara
Graduate School of System Informatics
Kobe University, Japan
1-1, Rokkodai, Nada, Kobe 657-8501
uehara@kobe-u.ac.jp

ABSTRACT

This paper introduces a method for *Multimedia Event Detection* (MED). Given training videos for a certain event, a classifier is constructed to identify videos displaying it. In particular, the problems of the *weakly supervised setting* and the *unclear event structure* are addressed in this paper. The first issue is associated with the loosely annotated training videos that usually contain many irrelevant shots. The second one is the difficulty of assuming the event structure in advance, because videos can be created by arbitrary camera and editing techniques. To overcome these problems, a *Hidden Conditional Random Field* (HCRF) is used where hidden states work as an intermediate layer to discriminate between relevant and irrelevant shots to the event. In addition, the relation among hidden states characterises the event structure. Thus, the above problems are managed by optimising hidden states and their relation, so as to distinguish videos where the event occurs from the rest of videos. Experimental results on TRECVID video data validate the effectiveness of HCRFs in this context.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation, Performance

Keywords

Multimedia Event Detection, Hidden Conditional Random Fields, Concept Detection, TRECVID

1. INTRODUCTION

With the explosive growth of video data on the Web, it is necessary to develop methods which analyse a large number of videos based on automatically extractable features, and accurately retrieve the ones of interest. This paper deals with the *Multimedia Event Detection* (MED) task to identify videos where a particular event occurs. An event is defined as a complex activity of objects at a specific place and time [19]. Figure 1 shows two videos where the event "birthday party" occurs. Here, *Video 1* is made of a single shot that continuously follows persons' actions, while *Video 2* is created by concatenating shots each shows a different scene from an isolated camera position. In contrast to the traditional task of retrieving shots with certain meanings, events in MED may be displayed in single shots like *Video 1* in Figure 1, or over shot sequences like *Video 2*. Thus, MED is more challenging than the traditional shot retrieval, because it requires to consider not only shots in a video but also their relation. This paper addresses two problems in MED described below.

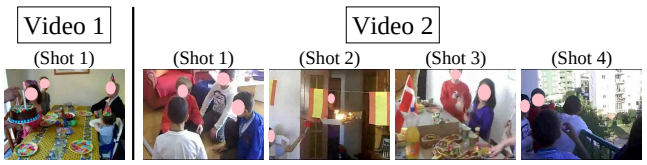

Figure 1: Example videos containing the event "birthday party".

(1) Weakly supervised setting: Training videos annotated with the occurrence or absence of a certain event are required to build a classifier distinguishing videos showing this event from the remaining videos. A video involves the time dimension where semantic meanings continuously change as video frames and audio samples are played sequentially. Due to the limited manpower and the subjectivity, it is impractical to manually annotate which segments in videos are relevant or irrelevant to an event. Thus, a classifier has to be built under weakly supervised setting, where each training video is loosely annotated to only indicate if the event is contained or not.

For the simplicity, videos annotated with an event's occurrence and its absence are called *positive videos* and *negative videos*, respectively. For example, *Video 1* and *Video 2* in Figure 1 are positive videos for the event "birthday party". However, as seen from *Shot 1* and *Shot 4* in *Video 2*, positive videos often contain many shots that are irrelevant to the event. Hence, in weakly supervised setting, the classifier construction process needs to identify what kind of shots are

ICMR '14, April 01 - 04 2014, Glasgow, United Kingdom

relevant or irrelevant to the event.

(2) Unclear event structure: An event consists of several sub-events. For instance, the event "birthday party" includes sub-events such as "blowing candle fires", "opening a gift", or "eating a birthday cake". To capture this kind of event structure, researchers traditionally limit the domain of videos. For example, in baseball videos, the event "home run" is presented by a shot sequence, where the first shot is taken behind the pitcher, the second shot follows the ball, and the third shot shows the batter running [3]. In addition, in movies, the conversation event is presented by a shot sequence, where shots showing one person and those presenting another one are repeated one after another [26]. Thus, events in the above videos can be easily detected based on heuristics which is implemented using pre-defined models, such as Hidden Markov Model (HMM) [3] or Finite State Machines (FSMs) [26].

Compared to this, we target real-world videos which are 'uncontrolled' in the sense that shots can be taken by arbitrary camera techniques and in arbitrary shooting environments, and what is more, they can be concatenated by arbitrary editing techniques. For such uncontrolled videos, the structure of an event cannot be assumed in advance, because relevant shots are characterised by significantly different features and their temporal relationship is unclear. Hence, by analysing training videos, we need to statistically mine features that are useful for characterising relevant shots, and their temporal relation.

To deal with the above weakly supervised setting and unclear event structure, we use a *Hidden Conditional Random Field* (HCRF) which is a probabilistic discriminative classifier with a set of hidden states [14]. We indirectly associate videos with an event by using hidden states as an intermediate layer, which characterises shots relevant or irrelevant to the event and their temporal relationship. In the HCRF, each shot in every positive or negative video is assigned to a hidden state which is characterised by a certain feature combination as well as the relevance to the event. Through this assignment, hidden states and their relation are optimised so as to discriminate between positive and negative videos. Experimental results show that, compared to the direct association between videos and events, the indirect association by HCRFs leads to more accurate performance. In addition, we intensively investigate several characteristics of HCRFs and indicate future directions of how to further improve the performance.

2. RELATED WORK

MED is one of tasks established in TRECVID which is an annual worldwide competition on video analysis and retrieval [19]. MED started from TRECVID 2011 and many methods have been developed so far [15, 8, 12, 9]. However, most of them adopt the same framework as the traditional shot retrieval. Specifically, despite the fact that each video consists of a different number of shots, features extracted from shots are aggregated into a vector with a fixed dimensionality. Based on this 'video-level' vector representation, classifiers used for shot retrieval (typically Support Vector Machine (SVM)) are built. Usually, a video-level vector is obtained by *max-pooling* [8, 12] or *average-pooling* [15], which takes the maximum or the average feature value across shots. In addition, *feature-accumulation* accumulates features extracted from various spatio-temporal regions in a video and represents their distribution with a probabilistic model [9]. The above video-level vectors are clearly too coarse. The reason is that max-pooling may overestimate values on features which are irrelevant to an event, and average-pooling and feature-accumulation may underestimate the ones on relevant features. Moreover, none of these approaches consider the temporal relation among shots. Compared to them, our method precisely models features in shots and their temporal relation using the intermediate layer of hidden states. We experimentally show that our approach outperforms max-pooling and average-pooling.

To model event structures, HMMs (or other types of generative models) have been traditionally used [3, 5]. However, HMMs are 'one-class' classifiers which only maximise the likelihood of positive videos without taking into account negative videos. In other words, the boundary between videos where an event occurs and the irrelevant videos is supported only from the positive side. Thus, without any prior knowledge, HMMs require a large number of positive videos to achieve accurate performance. Since events are very specific concepts, collecting many positive videos is difficult. Compared to HMMs, HCRFs are discriminative classifiers using both positive and negative videos. Many publications report that discriminative classifiers are considerably superior to one-class classifiers like HMMs [11, 25].

Furthermore, HMMs model hidden states using isolated probability distributions. This imposes the rigid restriction on HMMs, where due to the computational tractability, each shot needs to be regarded as conditionally-independent of the other shots. Here, a hidden state at each shot is chosen only by considering states and their transitions at the previous shot. Thus, HMMs cannot consider long-range dependencies among shots, which is undesirable for targeting events with unclear structures. In contrast, HCRFs model the conditional probability of the entire sequence using a single probability distribution, where features characterising long-range dependencies can be easily incorporated. In Section 4, we investigate the temporal characteristic of unclear event structures.

Existing methods which are the most related to ours are event detection using Conditional Random Fields (CRFs) [23, 24]. A CRF, which forms the basis of an HCRF, is a probabilistic discriminative classifier for labelling elements in a sequence [10]. Wang *et al.* used a CRF to label whether each shot in a video shows a highlight or not [23]. Also, targeting a network consisting of many sensors, Yin *et al.* used a CRF to annotate whether the recording of each sensor at every time point indicates the occurrence of an event [24]. Although CRFs can extract unclear event structures by treating long-range dependencies among shots, they require training videos where each shot is annotated with an event's occurrence or absence, thus cannot be used in weakly supervised setting. In contrast, HCRFs can handle this setting where CRF's shot labelling is performed on hidden states, and labelling results are combined to estimate the label for the whole video.

Finally, HCRFs have been successfully used in different applications such as object classification [14], action (gesture) recognition [27, 22], and audio analysis [7]. But, to the best of our knowledge, this paper describes the first application of HCRFs to MED.

3. EVENT DETECTION METHOD

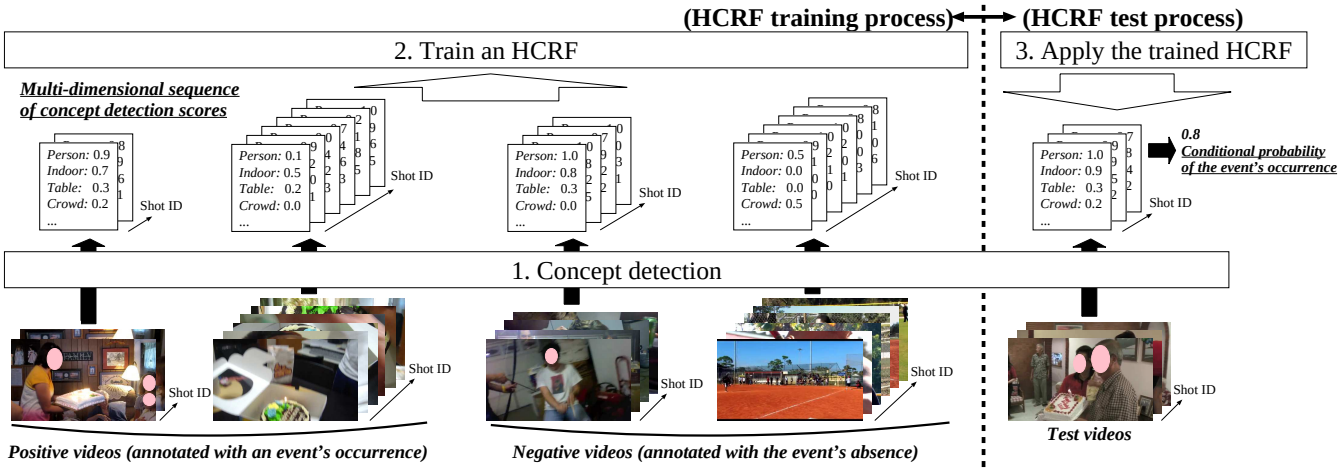

Figure 2: An overview of our MED method for "birthday party" used as an example event.

An event is 'highly-abstracted' in the sense that various objects interact with each other in different situations. Low-level features like colour, edges, and motion turned out to be ineffective in this context, because the set of shots relevant to a certain event has got a huge variance in the feature space. Hence, we adopt a *concept-based* approach which represents shots based on detection results of concepts such as *Person*, *Building*, or *Car* [21]. Since the detector of each concept is built using a large amount of training shots, it can be robustly detected irrespective of sizes, positions and directions in video frames. Thus, concept detection results can be considered as 'high-level' features, where the variation of relevant shots becomes smaller and can be modelled more easily.

Figure 2 shows an overview of our concept-based MED method. First, each video is divided into shots using a simple method detecting a shot boundary as a significant difference of colour histograms between two consecutive video frames. Then, concepts in each shot are detected. This yields *detection scores* representing the probability of a concept's presence. Thereby, as shown in the middle of Figure 2, every video is represented as a *multi-dimensional sequence* where each shot is temporally ordered and represented as a vector of concept detection scores.

Subsequently, given an event, an HCRF is trained using positive and negative videos, where the event's occurrence or absence is annotated at the video level (weakly supervised setting). Hidden states in the HCRF are probabilistically optimised so as to characterise features of shots relevant (or irrelevant) to the event and their temporal relation. Afterwards, as shown in the rightmost of Figure 2, the trained HCRF is used to compute the conditional probability of the event's occurrence for each test video. Finally, the sorted list of test videos based on these conditional probabilities is returned as an MED result. Below, we describe the concept detection process and the HCRF training/test process.

3.1 Concept Detection

The vocabulary of concepts should be sufficiently rich for describing various events. We use *Large-Scale Concept Ontology for Multimedia* (LSCOM) which is one of the most popular ontologies in the field of multimedia retrieval [13]. LSCOM defines a standardised set of 1,000 concepts that are selected based on their 'utility' for classifying content in videos, their 'coverage' for responding to a variety of queries, their 'feasibility' for automatic detection, and the 'availability' (observability) of large-scale training data. Our MED method characterises events using appearances of LSCOM concepts in shots.

It should be noted that LSCOM may contain no 'specific' concept to some events. For example, although *Birthday_Cake* and *Candle* are very specific to the event "birthday party", they are not defined in LSCOM. Ideally, for any event, all specific concepts should be defined because they are useful for detecting it. Nonetheless, without using specific concepts, events can be detected using related concepts. For example, if *Indoor*, *Food*, *Table*, and *Explosion_Fire* are shown in a shot, this shot probably contains also the concepts of *Birthday_Cake* and *Candle*. Our main purpose is not to build a large vocabulary of concepts, but to examine the effectiveness of HCRFs for MED, so we use LSCOM. The performance may be improved using a larger concept vocabulary like ImageNet [6] than LSCOM.

To build accurate concept detectors, a large number of training shots are required to cover diverse concept appearances. In addition, since a concept does not necessarily appear in all video frames in a shot, features need to be extracted from many video frames. To this end, we use the fast SVM training/test method and the fast feature extraction method based on matrix operation [18]. The former realises batch computation of similarities among many training shots, and the latter computes probability densities of many descriptors in a batch. These methods make SVM training/test and feature extraction about 10-37 and 5-7 times faster than the normal implementation, respectively.

Owing to the above fast methods, concept detection is conducted as follows: First, to characterise shapes of local regions, Scale-Invariant Feature Transform (SIFT) descriptors are extracted from every other frame. Then, hundreds of thousands of SIFT descriptors extracted from a shot are organised into the *GMM-SuperVector* (GMM-SV) representation, which represents their distribution using a Gaussian Mixture Model (GMM). Finally, for each concept, an SVM is constructed as a concept detector using 30,000 training shots. In total, detectors of 351 concepts are built because training data collected by the system [4] contain more than

one shot annotated as positive (concept presence).

3.2 Event Detection with HCRFs

Figure 3 illustrates an overview of an HCRF. Assume that a video $\boldsymbol{x}$ is represented as a multi-dimensional sequence of M concept detection scores, that is, if $\boldsymbol{x}$ has S shots, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_S)^{\mathrm{T}}$, where the i-th shot $\boldsymbol{x}_i$ $(1 \leq i \leq S)$ is represented as an M-dimensional vector $(x_{i,1}, \ldots, x_{i,M})^{\mathrm{T}}$, and $x_{i,c}$ $(1 \leq c \leq M)$ represents the c-th concept detection score of $\boldsymbol{x}_i$. Figure 3 depicts how to determine the event label $y \in \{0, 1\}$ of $\boldsymbol{x}$, where 0 and 1 mean the event's absence and occurrence, respectively. First, $\boldsymbol{x}_i$ is assigned to a hidden state $h_i \in \mathcal{H}$ from a set of all hidden states $\mathcal{H}$. Then, y is determined based on the sequence of hidden states $\boldsymbol{h} = (h_1, \cdots, h_S)^{\mathrm{T}}$ assigned to $\boldsymbol{x}$. More concretely, weakly supervised setting with y being loosely associated with the whole sequence $\boldsymbol{x}$ is managed by h_i which precisely examines the suitability of $\boldsymbol{x}_i$ for y. In addition, the structure of the event is characterised by the likelihood of transitions among hidden states in $\mathcal{H}$. The suitability of $\boldsymbol{x}$ for this structure is evaluated by the transition between h_i and h_{i+1}, assigned to $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$ respectively.

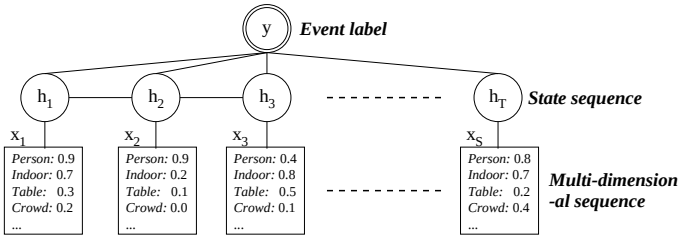

Figure 3: An illustration of our HCRF model.

The above HCRF is implemented based on the following conditional probability of y given $\boldsymbol{x}$:

$$P(y|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{\forall \boldsymbol{h} \in \mathcal{H}} P(y, \boldsymbol{h}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\sum_{\forall \boldsymbol{h} \in \mathcal{H}} e^{\Psi(y, \boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\theta})}}{\sum_{\forall y' \in \mathcal{Y}; \forall \boldsymbol{h} \in \mathcal{H}} e^{\Psi(y', \boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\theta})}}, \tag{1}$$

where the middle term indicates that $\boldsymbol{h}$ is marginalised out by taking the sum of $P(y, \boldsymbol{h}|\boldsymbol{x}, \boldsymbol{\theta})$ over all possible instances of $\boldsymbol{h}$ (i.e., all possible assignments of hidden states to $\boldsymbol{x}$). The middle term is further transformed into the rightmost one, where the numerator with the fixed y is normalised by the denominator taking the sum of numerators with all $y' \in \mathcal{Y}$ $(= \{0, 1\})$. Thus, the rightmost term can be considered as a conditional probability. Regarding the computation, for sequentially connected hidden states like the ones in Figure 3, the numerator and denominator can be efficiently computed by the 'brief propagation' algorithm, which propagates the intermediate result for each hidden state at $\boldsymbol{x}_i$ in both the backward and the forward directions [14].

Also, $\Psi(y, \boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\theta})$ in Equation (1) is called a *potential function* and parametrised by $\boldsymbol{\theta}$ as follows:

$$\begin{aligned} \Psi(y, \boldsymbol{h}, \boldsymbol{x}; \boldsymbol{\theta}) &= \sum_{i=1}^{S} \left(\boldsymbol{x}_i \cdot \boldsymbol{\theta}_{\mathrm{state}}(h_i) + \theta_{\mathrm{label}}(y, h_i) \right) \\ &+ \sum_{i=2}^{S} \theta_{\mathrm{trans}}(y, h_{i-1}, h_i) \quad , \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}$ consists of the following different types of parameters. For h_i assigned to $\boldsymbol{x}_i$, $\boldsymbol{\theta}_{\mathrm{state}}(h_i)$ represents an M-dimensional 'weight vector' where each dimension indicates the weight of a concept. Thus, the product $\boldsymbol{x}_i \cdot \boldsymbol{\theta}_{\mathrm{state}}(h_i)$ represents the degree of matching between $\boldsymbol{x}_i$ and h_i. In addition, h_i is associated with the 'label relevance' $\theta_{\mathrm{label}}(y, h_i)$, representing the relevance of h_i to the label y. Hence, the first term $(\boldsymbol{x}_i \cdot \boldsymbol{\theta}_{\mathrm{state}}(h_i) + \theta_{\mathrm{label}}(y, h_i))$ handles weakly supervised setting by examining whether each shot is relevant (or irrelevant) to the event. Furthermore, regarding the transition from h_{i-1} to h_i, $\theta_{\mathrm{trans}}(y, h_{i-1}, h_i)$ represents its relevance to y. Such a 'transition relevance' for each pair of hidden states characterises the temporal structure of the event. According to the above formulation, the management of weakly supervised setting and the extraction of an unclear event structure are reduced to the estimation of $\boldsymbol{\theta}$. In total, $\boldsymbol{\theta}$ consists of the set of $|\mathcal{H}|$ weight vectors $\boldsymbol{\theta}_{\mathrm{state}} = \{\boldsymbol{\theta}_{\mathrm{state}}(1), \cdots, \boldsymbol{\theta}_{\mathrm{state}}(|\mathcal{H}|)\}$, the set of $|\mathcal{Y}| \times |\mathcal{H}|$ label relevances $\boldsymbol{\theta}_{\mathrm{label}} = \{\theta_{\mathrm{label}}(y = 0, 1), \cdots, \theta_{\mathrm{label}}(y = 1, |\mathcal{H}|)\}$, and the set of $|\mathcal{Y}| \times |\mathcal{H}|^2$ transition relevances $\boldsymbol{\theta}_{\mathrm{trans}} = \{\theta_{\mathrm{trans}}(y = 0, 1, 1), \cdots, \theta_{\mathrm{trans}}(y = 1, |\mathcal{H}|, |\mathcal{H}|)\}$.

Suppose N training videos where the j-th training video $\boldsymbol{x}^{(j)}$ $(1 \leq j \leq N)$ consists of S_j shots, that is, $\boldsymbol{x}^{(j)} = (\boldsymbol{x}_1^{(j)}, \ldots, \boldsymbol{x}_{S_j}^{(j)})^{\mathrm{T}}$. In addition, $\boldsymbol{x}^{(j)}$ is annotated with the event label $y^{(j)} = 1$ if it is positive, otherwise $y^{(j)} = 0$. We estimate $\boldsymbol{\theta}$ which maximises the following log-likelihood based on conditional probabilities for $\boldsymbol{x}^{(j)}$ and $y^{(j)}$:

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{N} \log P(y^{(j)}|\boldsymbol{x}^{(j)}, \boldsymbol{\theta}) - \frac{||\boldsymbol{\theta}||^2}{2\sigma^2} \quad , \tag{3}$$

where the second term is the L2 regularisation term and useful for preventing $\boldsymbol{\theta}$ from being overfit to training videos. A smaller σ works as a stronger constraint which inhibits parameters in $\boldsymbol{\theta}$ to be extremely large. The optimal $\boldsymbol{\theta}^*$ is estimated by a gradient ascent method based on the derivative of Equation (3) in terms of each parameter in $\boldsymbol{\theta}$ [14]. Owing to the brief propagation algorithm, this derivative can be efficiently computed.

After $\boldsymbol{\theta}^*$ is obtained, the relevance score of each test video $\boldsymbol{x}$ to the event is computed as the conditional probability of $y = 1$ for $\boldsymbol{x}$, that is, $P(y = 1|\boldsymbol{x}, \boldsymbol{\theta}^*)$ based on Equation (1). The sorted list of test videos in terms of their relevance scores is returned as the MED result.

4. EXPERIMENTAL RESULTS

Our MED method has been tested using three datasets, *EV* consisting of 5,472 videos (51,857 shots), *BG* consisting of 4,992 videos (32,384 shots), and *TE* consisting of 27,033 videos (180,219 shots). According to the official instruction of the TRECVID MED task, for each of the 10 events, 100 positive videos have been selected from *EV*. Table 1 summarises these 10 events with the first, second, and third columns representing the event ID, its description, and the average number of shots contained in positive videos, respectively. Note that this average is not the average number of shots displaying an event. Due to weakly supervised setting, depending on positive videos, the event may be displayed in a single shot, in some shots, or in all shots.

The set of negative videos has been created by combining all videos in *BG* with some 'near-miss' videos in *EV* which are visually similar to positive videos, but do not contain

Table 1: Events addressed in our experiments.

Event ID	Event Description	Avr. # of shots
E006	Birthday party	10.69
E007	Changing a vehicle tire	10.32
E008	Flash mob gathering	25.12
E009	Getting a vehicle unstuck	5.38
E010	Grooming an animal	5.10
E011	Making a sandwich	14.06
E012	Parade	9.34
E013	Parkour	20.06
E014	Repairing an appliance	10.72
E015	Working on a sewing project	9.51

the event. Figure 4 depicts two examples of near-miss videos for the event "birthday party", where *Video 1* only shows a cake, and *Video 2* shows a man cooking cakes for a party. We firstly assumed that the performance will get degraded using near-miss videos as negative videos. The reason is that many test videos where an event occurs may be missed, because they may be similar to near-miss videos. However, our preliminary experiment has shown no significant difference of performance between using near-miss videos and not-using them. Hence, we have included near-miss videos into the set of negative videos.

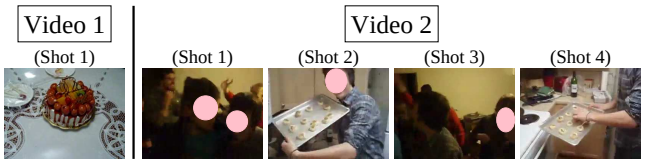

Figure 4: Examples of near-miss videos for "birthday party".

HCRFs built using positive and negative videos described above have been tested on videos in *TE*. Each result has been evaluated by an Average Precision (AP), which is the average of precisions at positions of test videos where an event occurs. A larger AP means that such test videos are ranked at higher positions.

Finally, it should be noted that each HCRF has been trained using 100 positive videos and more than $4,000$ negative videos. This setting may cause the imbalanced problem which makes it difficult to build a well-generalised HCRF [2]. When the number of negative videos (majority class) is much higher than that of positive ones (minority class), the meaningless HCRF which classifies all videos as negative may be regarded as accurate on training videos.

Table 2 shows a preliminary experiment on the imbalanced problem. Targeting three events *E006*, *E009* and *E013*, this table presents the comparison of APs between *All* using all negative videos and *Sampled* using randomly sampled $1,000$ negative videos. In particular, the upper line of *Sampled* represents the average and standard deviation of APs, obtained by 10 runs using different sets of negative videos. The lower line presents the maximum and minimum APs in the above 10 runs. As shown in Table 2, by comparing APs of *All* and the average APs of *Sampled*, the former seems to be slightly superior to the latter. Thus, the imbalanced problem has no strong influence on the performance of HCRFs. In addition, although *Sampled* sometimes outperforms *All*, its APs are considerably varied. To avoid this variance and explicitly evaluate the effectiveness of HCRFs, the following experiments have been conducted using all negative videos.

Table 2: Performance comparison between *All* and *Sampled*.

Event ID	*E006*	*E009*	*E013*
All	0.062	0.128	0.174
Sampled	0.079 ± 0.006	0.094 ± 0.015	0.151 ± 0.013
(max, min)	$(0.090, 0.069)$	$(0.119, 0.064)$	$(0.179, 0.135)$

4.1 Parameter Initialisation

Since the objective function in Equation (3) has many local maxima, setting a proper initial $\boldsymbol{\theta}$ is crucial for building an HCRF with an effective $\boldsymbol{\theta}^*$. For this, we borrow the idea of initialisation used in HMMs [16]. First, an initial $\boldsymbol{\theta}$ is determined based on the 'hard-assignment' of shots to hidden states. Here, $\boldsymbol{\theta}$ is initialised only using the maximum likelihood sequence of hidden states for each training video. Then, the initial $\boldsymbol{\theta}$ is refined to $\boldsymbol{\theta}^*$ by the 'soft-assignment' where all possible sequences of hidden states are considered based on Equation (1). Our method for $\boldsymbol{\theta}$ initialisation is summarised below.

Since hidden states are shared by all shots, it is reasonable to initialise $\boldsymbol{\theta}_{\text{state}}$ so as to characterise their distribution. To this end, shots are grouped into the same number of clusters to that of hidden states. Because training videos for each event contain more than $32,000$ shots, a fast clustering method [1] is used. However, it is not reasonable to initialise $\boldsymbol{\theta}_{\text{state}}$ using cluster centres. Since the range of concept detection scores is between 0 and 1, no cluster centres take values less than 0. On the other hand, values of weight vectors in $\boldsymbol{\theta}_{\text{state}}$ can be negative. Hence, an initial $\boldsymbol{\theta}_{\text{state}}$ is required not only to reflect the clustering result, but also to take positive and negative values.

To obtain a proper initial $\boldsymbol{\theta}_{\text{state}}$, we construct a CRF that is a probabilistic model having a similar structure to HCRFs [10]. As shown in Figure 5, the CRF is made by removing the event label layer of an HCRF. The CRF has parameters corresponding to $\boldsymbol{\theta}_{\text{state}}$ in the HCRF. These are optimised by a similar probabilistic optimisation to the HCRF, concretely, by focusing on the relation between each shot and its 'observable' state label. Thus, when the CRF is optimised by regarding the cluster index of each shot as the state label, parameters corresponding to $\boldsymbol{\theta}_{\text{state}}$ can reflect the clustering result by taking some negative values, and are used as the initial ones. In contrast to HCRFs, optimised parameters of the CRF are guaranteed as global optimum [10].

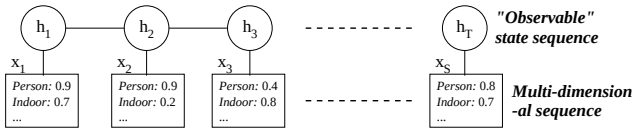

Figure 5: An illustration of a CRF.

In addition, using the CRF, the maximum likelihood sequence of state labels for each training video is computed and used as a sequence of hidden states. By only considering such sequences, initial $\boldsymbol{\theta}_{\text{label}}$ and $\boldsymbol{\theta}_{\text{trans}}$ are obtained

based on Equation (3). In this case, Equation (3) becomes convex [27], and initial $\boldsymbol{\theta}_{\text{label}}$ and $\boldsymbol{\theta}_{\text{trans}}$ can be globally optimised by a conventional gradient ascend method. Our preliminary experiment on three events *E006*, *E009* and *E013* shows that, compared to random initial values of $\boldsymbol{\theta}$, the above initialisation improves the objective function value of Equation (3) by 5.0 to 20.8%.

4.2 Evaluation for Weakly Supervised Setting

To examine the effectiveness of HCRFs for weakly supervised setting, we have compared three methods, SVM_{avr}, SVM_{max}, and *HCRF*. As seen from Equation (2), hidden states use linear combinations of concept detection scores. Thus, SVM_{avr} constructs a linear SVM where the decision function linearly combines concept detection scores of a video-level vector obtained by average-pooling. Similarly, SVM_{max} constructs a linear SVM based on video-level vectors by max-pooling. The SVM parameter for penalising mis-classified training videos has been heuristically set to 2.

By comparing SVM_{avr} and SVM_{max} to *HCRF*, we aim to reveal the effectiveness of *HCRF*, where hidden states are used as an intermediate layer to precisely characterise event structures based on shot-level vectors of concept detection scores. In *HCRF*, 10 hidden states are used, and due to the computational cost the maximum number of iterations for estimating $\boldsymbol{\theta}^*$ is set to 50. Note that the performance of *HCRF* significantly depends on the parameter σ for L2 regularisation. We have tested each $\sigma \in \{0.5, 1, 2, 4\}$ and have manually selected the one achieving the best performance. This manual selection aims to avoid underestimating the performance of *HCRF*. One solution for the σ selection problem will be discussed later.

Figure 6 shows the performance comparison between SVM_{avr}, SVM_{max}, and *HCRF*. For each event listed in the vertical direction, APs are depicted in the horizontal direction where different marks are used depending on methods. The bottom entry shows the Mean of APs (MAPs) over all 10 events. As can be seen from Figure 6, for 8 of 10 events, *HCRF* outperforms SVM_{avr} and SVM_{max}. In particular, to check the significance of this result, we have conducted 'randamisation test' [20]. It randomly swaps APs of *HCRF* and those of SVM_{avr} (or SVM_{max}) by assuming that there is no significant performance difference (null hypothesis), and examines whether the difference between their actual MAPs is statistically unlikely or not. As a result, we have confirmed that *HCRF* is superior to SVM_{avr} and SVM_{max} with the significance level of 3%. This validates that hidden states of HCRFs work well to discriminate between relevant and irrelevant shots to an event.

4.3 Evaluation for Extracting Unclear Event Structures

We investigate how useful hidden states are for characterising unclear event structures. First, we compare performances of HCRFs with different numbers of hidden states. This aims to examine whether a larger number of hidden states cover a more diversity of shots relevant (or irrelevant) to an event. Figure 7 shows the transition of APs depending on different numbers of hidden states. As shown in the horizontal axis, APs obtained by 3, 5, 10 and 15 hidden states are plotted, in terms of *E006*, *E009*, *E013* and the MAP on these events. As can be seen from Figure 7, although the performance improvement is relatively unclear

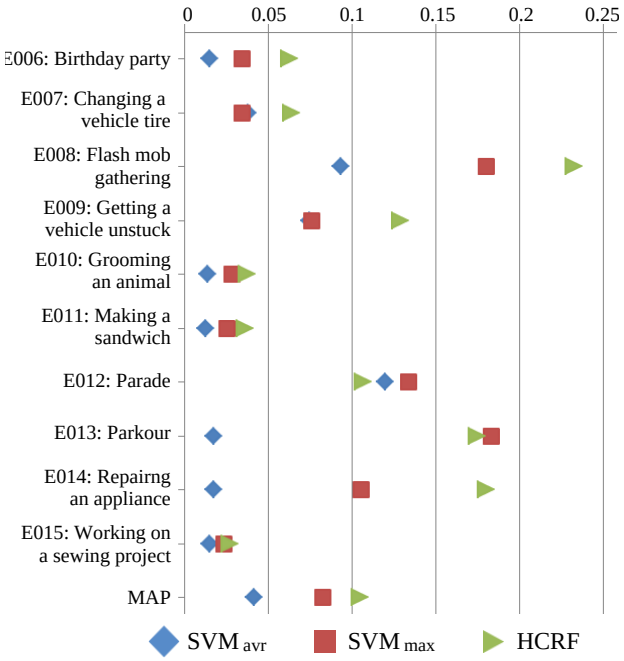

Figure 6: Performance comparison between SVM_{avr}, SVM_{max} and *HCRF*.

at the event level, the overall performance (i.e., MAP) is gradually improved using a larger number of hidden states. This indicates that a more diversity of shots can be appropriately covered using more hidden states. However, using 15 hidden states slightly improves the MAP using 10 hidden states. Thus, as the trade-off between the performance and the computational cost, using 10 hidden states can be considered as a reasonable choice.

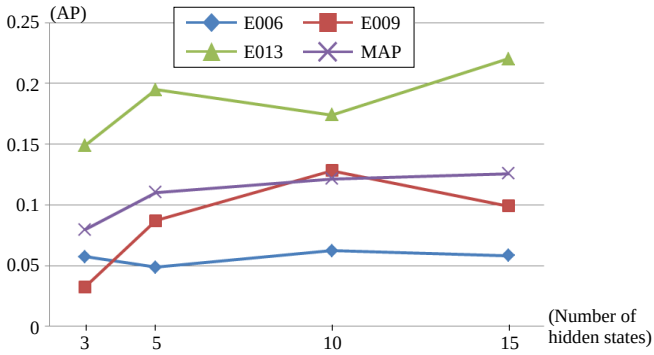

Figure 7: Performance transition using different numbers of hidden states.

Also, we checked which concepts are characterised by learnt hidden states. As a result, hidden states which are associated with large label relevances for an event's occurrence (i.e., $\theta_{\text{label}}(y = 1, h)$), are confirmed to appropriately characterise relevant concepts, such as *Food* and *Explosion_Fire* for *E006: Birthday party*, and *Vehicle* and *Rocky_Ground* for *E009: Getting a vehicle unstuck*. However, these hidden states also wrongly characterise several irrelevant concepts, like *Basketball* and *Rescue_Vehicle* for *E006*, and *Bar_Pub*

and *Airplane_Flying* for *E009*. One main reason is the current imperfect concept detection which only uses a single image feature (SIFT). Thus, we will incorporate motion and audio features into concept detection to improve its performance. This will lead to improve the MED performance.

Now, based on the flexibility of the potential function in Equation (2), we explore the temporal structures of events by comparing *HCRF* to $HCRF_{\text{no}}$ and $HCRF_{\text{win}}$. Regarding $HCRF_{\text{no}}$, one may think that there are no clear temporal structures of events in uncontrolled videos, so the transition among hidden states is meaningless and degrades the performance. Thus, $HCRF_{\text{no}}$ does not consider the transition among hidden states by removing the term $\sum \theta_{\text{trans}}(y, h_{i-1}, h_i)$ from Equation (2). Regarding $HCRF_{\text{win}}$, HCRFs can deal with long-range dependencies among shots. In particular, [22] reported that the performance of gesture recognition is improved using a 'window feature', which combines features at the previous, current and next time points. $HCRF_{\text{win}}$ implements such a window feature, where $\boldsymbol{x}_i$ of the i-shot in Equation (2) is the concatenation of concept detection scores at the $(i-1)$, i and $(i+1)$-th shots.

Table 3 shows the performance comparison among $HCRF_{\text{no}}$, *HCRF* and $HCRF_{\text{win}}$. First, *HCRF* significantly outperforms $HCRF_{\text{no}}$. This indicates that the transition among hidden states works well to characterise temporal structures of events. In particular, we found that many videos where an event does not occur are falsely detected, only because they contain shots which are similar to relevant shots to the event. For example, for *E006: Birthday party*, falsely detected videos just contain shots displaying children (many positive videos show birthday parties of children). In addition, for *E009: Getting a vehicle unstuck*, shots in falsely detected videos just show cars. Thus, the transition among hidden states can be considered as effective constraints to alleviate the above false-positive detection.

Table 3: Performance comparison among $HCRF_{\text{no}}$, *HCRF*, and $HCRF_{\text{win}}$.

	E006	*E009*	*E013*
$HCRF_{\text{no}}$	0.030	0.061	0.026
HCRF	0.062	0.128	0.174
$HCRF_{\text{win}}$	0.046	0.084	0.176

The comparison between *HCRF* and $HCRF_{\text{win}}$ shows that even if the window feature is used, the performance is similar or even degraded. This implies that each gesture addressed in [22] is taken by a single camera, so time points are continuous and have strong temporal correlation. On the other hand, each uncontrolled video consists of shots taken by different cameras. So, these shots are discontinuous, and their temporal correlation is often corrupted by inserting shots, which display a different meaning than those of surrounding shots. Therefore, we need to consider longer-range dependencies than $HCRF_{\text{win}}$ while flexibly treating the distorted order of shots.

To capture the above long-range dependencies, we plan to model the continuity of a concept's presence. Each concept can be considered to have some appearance patterns, related to the story of a video. For example, when the concept plays an important role, it is present in many shots, otherwise it is not present so often. Thus, using the time series segmentation method developed in [17], we will divide a video into shot sequences each of which is characterised by a probabilistically distinct pattern of the concept's presence. Such a pattern indicates the continuity of the concept's presence at each shot. Then, $\boldsymbol{x}_i$ in Equation (2) will be enlarged by adding dimensions each of which represents the continuity of a concept's presence at the i-th shot.

4.4 Bagging of HCRFs

We described two factors, σ (see Section 4.2) and a set of negative videos (see Table 2), which cause unstable results of HCRFs. By checking such results, we acquired one finding that, videos where an event occurs are ranked at relatively high positions, while videos where it does not occur are ranked at different positions. Thus, rather than cross validation to select an effective σ or set of negative videos, combining unstable results is expected to improve the performance. Therefore, in analogy with bagging which combines classification results obtained by different subsets of training videos, we combine results by different σs and different sets of negative videos into a single result.

We have devised two bagging approaches, $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$. For an event, $HCRF^{(\sigma)}_{\text{bag}}$ combines results obtained by four HCRFs, each of which is built using $\sigma \in \{0.5, 1, 2, 4\}$ and the set of all negative videos. In $HCRF^{(\sigma,n)}_{\text{bag}}$, 40 HCRFs are combined where each one uses $\sigma \in \{0.5, 1, 2, 4\}$ and a set of randomly sampled $1,000$ negative videos. In both of $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$, for each test video $\boldsymbol{x}$, the sum of $P(y = 1|\boldsymbol{x}, \boldsymbol{\theta}^*)$s by different HCRFs, is used as the final relevance score to the event.

Table 4 shows the performance comparison between the above bagging approaches, and *HCRF* in Figure 6 where the best σ is manually selected. In Table 4, APs in bold font indicate that $HCRF^{(\sigma)}_{\text{bag}}$ or $HCRF^{(\sigma,n)}_{\text{bag}}$ outperforms *HCRF*. Overall, as seen from the column *MAP*, both of $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$ are more accurate than *HCRF*. Randomisation test indicates that $HCRF^{(\sigma,n)}_{\text{bag}}$ outperforms *HCRF* with the significance level of 8%, and there is no significant difference between $HCRF^{(\sigma)}_{\text{bag}}$ and *HCRF*. The important finding here is that bagging leads to results which are the same or superior to the one obtained by manual.

Furthermore, Table 4 presents different characteristics of $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$. Except *E009*, $HCRF^{(\sigma)}_{\text{bag}}$ yields great improvement on *E008*, *E013* and *E014*, while its performance is similar to *HCRF* on the other events. Thus, bagging with different σs and all negative videos, works quite well on some events. On the other hand, $HCRF^{(\sigma,n)}_{\text{bag}}$ provides modest improvement on most events, but it is significantly degraded on some events like *E008* and *E014*, due to insufficient negative videos used to build each HCRF. Hence, $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$ can be considered as complementary. One interesting research topic is how to select the best bagging strategy depending on events. If the best of $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$ could be correctly selected for each event in Table 4, the MAP would become 0.131.

5. CONCLUSION AND FUTURE WORK

In this paper, we addressed weakly supervised setting and unclear event structures, and introduced a method using HCRFs. In an HCRF, hidden states are used as an intermediate layer, which characterises shots relevant (or irrelevant) to an event, and their temporal relation. These hidden

Table 4: Performance comparison among *HCRF*, $HCRF^{(\sigma)}_{\text{bag}}$ and $HCRF^{(\sigma,n)}_{\text{bag}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*
HCRF	0.062	0.063	0.231	0.128	0.036	0.035
$HCRF^{(\sigma)}_{\text{bag}}$	**0.065**	0.050	**0.262**	0.072	0.032	**0.041**
$HCRF^{(\sigma,n)}_{\text{bag}}$	**0.096**	**0.083**	0.200	**0.137**	0.031	**0.060**

	E012	*E013*	*E014*	*E015*	MAP
HCRF	0.106	0.174	0.179	0.026	0.104
$HCRF^{(\sigma)}_{\text{bag}}$	0.098	**0.251**	**0.206**	0.023	**0.110**
$HCRF^{(\sigma,n)}_{\text{bag}}$	**0.160**	**0.220**	0.162	**0.029**	**0.118**

states are probabilistically optimised so as to discriminate between positive and negative videos. Experimental results showed that HCRFs are effective for weakly supervised setting, and reasonably capture unclear event structures.

In addition to future works described in Section 4, we will address the following two issues: The first one is that although each hidden state currently uses linear combination of concept detection scores in the potential function, its discriminative power is not so strong. Thus, we will replace linear combination with a kernel-like combination where for each shot in a video, its similarity to every shot in training videos is weighted and summed up. In other words, weights for such similarities are parameters to be optimised in an HCRF. However, this causes significant increase of parameters and requires expensive computational cost. Thus, the second issue is to parallelise the HCRF training process using several processors.

6. REFERENCES

[1] *bayon - a simple and fast clustering tool.* `http://code.google.com/p/bayon/`.

[2] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *Proc. of ECML 2004*, pages 39–50. 2004.

[3] R. Ando, K. Shinoda, S. Furui, and T. Mochizuki. Robust scene recognition using language models for scene contexts. In *Proc. of MIR 2006*, pages 99–106, 2006.

[4] S. Ayache and G. Quénot. Video corpus annotation using active learning. In *Proc. of ECIR 2008*, pages 187–198, 2008.

[5] M. Barnard and J. Odobez. Sports event recognition using layered HMMs. In *Proc. of ICME 2005*, pages 1150–1153, 2005.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. of CVPR 2009*, pages 248–255, 2009.

[7] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt. Hidden conditional random fields for phone classification. In *Proc. of INTERSPEECH 2005*, pages 1117–1120, 2005.

[8] H. Cheng *et al.* SRI-Sarnoff AURORA system at TRECVID 2012: Multimedia event detection and recounting. In *Proc. of TRECVID 2012*, 2012.

[9] N. Inoue, T. Wada, Y. Kamishima, K. Shinoda, and S. Sato. TokyoTech+Canon at TRECVID 2011. In *Proc. of TRECVID 2011*, 2011.

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML 2001*, pages 282–289, 2001.

[11] X. Li and C. G. Snoek. Visual categorization with negative examples for free. In *Proc. of MM 2009*, pages 661–664, 2009.

[12] J. Liu, S. McCloskey, and Y. Liu. Local expert forest of score fusion for video event classification. In *Proc. of ECCV 2012*, pages 397–410, 2012.

[13] M. Naphade *et al.* Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[14] A. Quattoni, S. Wang, L. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852, 2007.

[15] R. Aly *et al.* AXES at TRECVid 2012: KIS, INS, and MED. In *Proc. of TRECVID 2012*, 2012.

[16] S. Young et al. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2009.

[17] K. Shirahama and K. Uehara. A novel topic extraction method based on bursts in video streams. *International Journal of Hybrid Information Technology (IJHIT)*, 1(3):21–32, 2008.

[18] K. Shirahama and K. Uehara. Kobe university and Muroran institute of technology at TRECVID 2012 semantic indexing task. In *Proc. of TRECVID 2012*, 2012.

[19] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. of MIR 2006*, pages 321–330, 2006.

[20] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. of CIKM 2007*, pages 623–632, 2007.

[21] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.

[22] S. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Proc. of CVPR 2006*, pages 1521–1527, 2006.

[23] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong. Semantic event detection using conditional random fields. In *Proc. of CVPRW 2006*, page 109, 2006.

[24] J. Yin, D. H. Hu, and Q. Yang. Spatio-temporal event detection using dynamic conditional random fields. In *Proc. of IJCAI 2009*, pages 1321–1326, 2009.

[25] H. Yu, J. Han, and K.-C. Chang. PEBL: Web page classification without negative examples. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):70–81, 2004.

[26] Y. Zhai, Z. Rasheed, and M. Shah. A framework for semantic classification of scenes using finite state machines. In *Proc. of CIVR 2004*, pages 279–288. 2004.

[27] J. Zhang and S. Gong. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1):197 – 203, 2010.