Noname manuscript No.
(will be inserted by the editor)

Weakly Supervised Detection of Video Events Using Hidden Conditional Random Fields

Kimiaki Shirahama · Marcin Grzegorzek · Kuniaki Uehara

Received: date / Accepted: date

Abstract *Multimedia Event Detection* (MED) is the task to identify videos in which a certain event occurs. This paper addresses two problems in MED: *weakly supervised setting* and *unclear event structure*. The first indicates that since associations of shots with the event are laborious and incur annotator's subjectivity, training videos are loosely annotated as whether the event is contained or not. It is unknown which shots are relevant or irrelevant to the event. The second problem is the difficulty of assuming the event structure in advance, due to arbitrary camera and editing techniques. To tackle these problems, we propose a method using a *Hidden Conditional Random Field* (HCRF) which is a probabilistic discriminative classifier with a set of hidden states. We consider that the weakly supervised setting can be handled using hidden states as the intermediate layer to discriminate between relevant and irrelevant shots to the event. In addition, an unclear structure of the event can be exposed by features of each hidden state and its relation to the other states. Based on the above idea, we optimise hidden states and their relation so as to distinguish training videos containing the event from the others. Also, to exploit the full potential of HCRFs, we establish approaches for training video preparation, parameter initialisation and fusion of multiple HCRFs. Experimental results on TRECVID video data validate the effectiveness of our method.

Keywords Multimedia event detection · Hidden conditional random fields · Weakly supervised setting · Unclear event structure

K. Shirahama and M. Grzegorzek
Pattern Recognition Group, University of Siegen
Hoelderlinstr. 3, D-57076 Siegen, Germany
Tel.: +49 271 740 3972
Fax: +49 271 740 1 3972
E-mail: {kimiaki.shirahama,marcin.grzegorzek}@uni-siegen.de

Kuniaki Uehara
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 657-8501 Japan
Tel.: +81 78 803 6215
Fax: +81 78 803 6316
E-mail: uehara@kobe-u.ac.jp

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. As a result of the recent research progress, accurate retrieval has been possible in terms of some objects, actions and scenes (e.g., car, running and outdoor) [27]. However, these meanings are too general, and thus impractical, as to identify videos users want to watch. This paper deals with *Multimedia Event Detection* (MED) to identify videos in which a particular event occurs. The event is a high-level meaning defined as a complex activity of objects at a specific place and time [25]. Such events are much more useful for practical applications than general meanings addressed before.

Let us consider how an event is presented in videos. Figure 1 shows two videos where the event "birthday party" occurs. The simplest presentation of an event is to use a long take shot which continuously follows objects related to the event. For example, *Video 1* in Figure 1 is made of a single shot that tracks the whole series of bringing the birthday cake to the table and blowing out candles. However, the screen can only capture a spatially limited part of an event, and the time duration of a video is limited. Consequently, a single

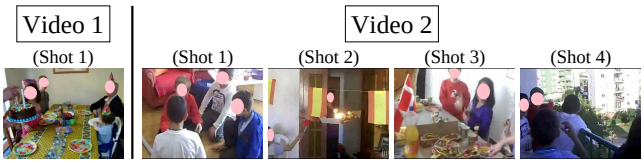


Fig. 1 Example videos containing the event "birthday party".

shot is often ineffective and inefficient to capture movements and interactions of multiple objects. Thus, editing is employed where shots showing key moments are connected to present the event in a compact form [3]. In Figure 1, *Video 2* consists of four shots which show chatting before the birthday party (*Shot 1*), bringing the cake (*Shot 2*), eating it (*Shot 3*) and playing after the party (*Shot 4*). Therefore, MED requires to consider shots in a video as well as their relation.

This paper addresses two problems in MED described below:

(1) Weakly supervised setting: MED can be formulated as a binary classification problem where a classifier is constructed to distinguish videos showing one event from the remaining ones. However, the classifier has to be constructed under the *weakly supervised setting* where each training video is annotated only with the occurrence or absence of the event, despite the fact that this video may include several semantically different shots. The weakly supervised setting arises due to the following two reasons: First, it is labour-intensive to annotate each shot contained in a video. Second, videos are 'continuous media' where semantic meanings are continuously conveyed as video frames and audio samples are played over time [10]. Because of this temporal continuity of meanings, any segment of a video can become a meaningful unit [30].

More concretely, humans usually tend to relate each shot in a video to surrounding ones. For example, in *Video 2* in Figure 1, there is no doubt that *Shot 2* and *3* show the birthday party. Based on this knowledge, one can deduce that *Shot 1* and *4* are related to the party, as chatting before and playing after "cake eating". This kind of shot relation makes it ambiguous to determine the boundary of an event. For *Video 2*, one may think that the birthday party is shown only in *Shot 2* and *3*, while someone else may think that it is shown during the whole of the video, by regarding *Shot 1* and *4* as parts of the party based on the shot relation described above. Thus, objective annotation is only possible at the video level in terms of whether each video contains the event or not. No annotation is provided for shots.

We call videos annotated with an event's occurrence and its absence *positive videos* and *negative videos*, respectively. For example, *Video 1* and *2* in Figure 1 are positive videos for the event "birthday party". It is notable that due to the weakly supervised setting, positive videos contain many shots that are irrelevant to the event. For example, *Shot 1* and *4* in *Video 2* are irrelevant because nothing related to the birthday party is displayed. Furthermore, shots similar to them are often observed in videos showing events other than "birthday party". Hence, to build an accurate classifier under the weakly supervised setting, we need to discriminate between relevant and irrelevant shots to an event.

(2) Unclear event structures: An event is a meaning specialised by the combination of objects, actions and scenes. In other words, the same object, action or scene is related to different events. Thus, relations among objects, actions and scenes are needed to specialise meanings presented in a video. We define the following two types of relations as *event structures*. The first type specialises meanings within a shot. For example, only with the appearance of food in a shot, it cannot be judged whether it is for eating or cooking. But, if the shot shows a dinning room, the food can be regarded as for eating. The second type of event structure specialises meanings over a shot sequence. This is attributed to editing that produces a new meaning by connecting multiple shots [3]. For example, the action of eating is common to many parties. But, if a shot showing a birthday cake is followed by a shot where a person eats something, it can be interpreted that these shots belong to a birthday party. Like this, event structures work as constraints to precisely examine the occurrence of an event in a video.

However, we target real-world videos created by non-professional users. Such videos are 'unconstrained' [14] where 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. As a result, appearances of objects, actions and scenes vary greatly, and shots are connected in different orders. Thus, event structures are 'unclear' in the sense that they cannot be assumed in advance. Hence, by analysing training videos, we need to statistically mine characteristic shots that are relevant (or irrelevant) to an event, and their temporal relation. It should be noted that we aim to extract event structures under the weakly supervised setting, where only the video level annotation is available and no shot is annotated. Thus, the extraction of event structures requires to solve the weakly supervised setting at the same time.

To jointly address these problems, we use a *Hidden Conditional Random Field* (HCRF) which is a probabilistic discriminative classifier with a set of hidden states [21]. For the weakly supervised setting, we indirectly associate a video with an event, using hidden

states as the intermediate layer to discriminate between relevant and irrelevant shots. In addition, each hidden state represents characteristic features and has the relation to the other states. We use such hidden states and their relation to characterise the structure of the event. Based on the above idea, we devise an HCRF which detects the event in a video by assigning each shot to a hidden state. This assignment is controlled by not only matching features of each shot with the ones of a hidden state, but also considering the relation (transitions) among hidden states. Then, the event's occurrence in the video is predicted by collecting relevance values of assigned hidden states. Thus, the weakly supervised setting and unclear event structure can be jointly handled by optimising hidden states (i.e., their features and relevance values) and the relation among them, so as to discriminate between positive and negative videos for the event.

Finally, we need to investigate several issues to train effective HCRFs. First, although negative videos need to cover various kinds of videos, it is unknown whether an HCRF can be appropriately trained on the biased set of training videos, where the number of negative videos is much larger than that of positive ones. We experimentally show that the performance of HCRFs is stable in terms of the number of negative videos (even through some of them are very similar to positive videos). Second, since the optimisation of an HCRF has many local maxima, it is necessary to properly initialise parameters which define hidden states and their relation. We develop a parameter initialisation method based on the distribution of shots and their connections in training videos. Third, the performance of HCRFs is unstable depending on the hyperparameter (regularisation parameter). Instead of choosing the best one, we devise a method which improves the performance by fusing unstable results. Based on the above investigation, we validate the effectiveness of HCRFs for the weakly supervised setting and unclear event structures.

2 Related Work

MED is one of the tasks established in TRECVID which is an annual worldwide competition on video analysis and retrieval [25]. MED started with TRECVID 2011 and many methods have been developed so far [1, 7, 13, 18]. However, most of them adopt the same approach as traditional shot retrieval. With respect to this, a shot is a basic unit which captures coherent meanings, in the sense that these are spatially and temporally continuous [8]. On the other hand, a video is a sequence of shots containing varied meanings. Despite this intrinsic difference between a shot and a video, most of the existing methods employ the traditional shot retrieval approach. Here, features extracted from shots in a video are aggregated into a 'video-level' vector, which represents overall meanings in the video. Based on this, classifiers used for traditional shot retrieval (typically Support Vector Machine (SVM)) are built.

A video-level vector is usually obtained by *max-pooling* [7, 18] or *average-pooling* [1], which takes the maximum or the average feature value over shots in a video. In addition, *feature-accumulation* accumulates features extracted from various spatio-temporal regions in a video, and creates a vector which represents the probability distribution of these features [13]. However, video-level vectors are clearly too coarse, because max-pooling may over-estimate values on features which are irrelevant to an event, and average-pooling and feature-accumulation may under-estimate the ones on relevant features. Moreover, none of them can consider the temporal relation among shots. In contrast to this view and according to the sequential nature of videos, we represent each video as a sequence of vectors each expressing a shot. Then, using an HCRF, training videos are abstracted into hidden states, which represent characteristic features of shots relevant (or irrelevant) to an event. And, the relation among hidden states captures representative shot transitions in the event. Thus, our method carries out more logical and precise modelling of events than existing methods [1, 7, 13, 18]. We experimentally show that our method outperforms max-pooling and average-pooling.

Recently, some researchers have proposed methods which discriminate between relevant and irrelevant shots (segments) to an event based on the latent SVM model [16, 31]. A video is associated with latent (binary) variables, each of which indicates whether a shot is used to compute the decision function. A latent SVM is trained by iterating the following two processes: The first one finds the configuration of latent variables (i.e., selection of shots) which best match the current decision function, and the other updates parameters of this function by assuming that latent variables for each training video are fixed. Notice that this method can handle the weakly supervised setting but cannot extract event structures, because it just selects relevant shots. Compared to this, both of these problems can be managed by HCRFs where hidden states show abstract representations of relevant (or irrelevant) shots. Also, the method in [29] uses Fisher kernel encoding to characterise feature transitions over shot sequences in an event. However, this cannot extract features representing characteristic shots for the event. In contrast, HCRFs can extract event structures describing both of characteristic features within shots and their transitions.

Apart from MED, event structures are traditionally captured by limiting 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 [2]. In movies, the conversation event is presented by a shot sequence, where shots showing one person and those showing another one are repeated one after another [37]. Thus, events can be easily detected based on the above heuristics which is implemented using pre-defined models, such as Hidden Markov Models (HMMs) [2] or Finite State Machines (FSMs) [37]. Compared to this, we target unconstrained videos and aim to extract event structures that cannot be assumed in advance.

In addition, HMMs (and other generative models) [2, 5] are 'one-class' classifiers which only maximise the likelihood of positive videos without taking negative videos into account. This means that the boundary between videos in which an event occurs and the irrelevant videos is formed only by positive videos. In other words, HMMs merely extract the region where positive videos are densely populated. Thus, without any heuristics described above, HMMs require a large number of positive videos to accurately define the region that covers videos containing the event. Since an event is a highly-specialised meaning, videos containing it are rare, so collecting many positive videos is difficult. Compared to HMMs, HCRFs are 'two-class' classifiers which maximise the discrimination between positive and negative videos. This enables us to effectively define the boundary between videos containing the event and the others. Many publications report that two-class classifiers are considerably superior to one-class classifiers like HMMs [17, 36].

Our preliminary experiment showed that the performance of HMMs built on positive videos is quite poor, and significantly improved by additionally using HMMs built on negative videos. Since negative videos include all kinds of videos except for positive ones, an HMM built on them works as a prior distribution representing how features in shots and shot transitions are distributed in the general case. Thus, we can examine how much a video is biased to the region where positive videos are populated, by computing the difference between the probability obtained by an HMM built on positive videos and the one by the HMM on negative ones. A similar approach is popularly used in speaker adaptation in audio recognition [35]. Even with the above improvement, in Section 4.2, we demonstrate that HCRFs significantly outperform HMMs.

Furthermore, although HMMs can extract event structures represented by hidden states, they have no mechanism to discriminate between relevant and irrelevant shots to an event, that is, all shots in positive videos are assumed as relevant. Thus, HMMs are not suitable for the weakly supervised setting, because extracted event structures are significantly affected by irrelevant shots contained in positive videos. In contrast, HCRFs can characterise event structures by discriminating between relevant and irrelevant shots. In Section 4.3, we exhibit that, even in the weakly supervised setting, meaningful event structures can be extracted by HCRFs.

Existing methods that are closely related to ours are event detection using Conditional Random Fields (CRFs) [33, 34]. A CRF, which forms the basis of an HCRF, is a probabilistic discriminative classifier for labelling elements in a sequence [15]. Wang *et al.* used a CRF to label whether each shot in a video shows a highlight or not [33]. 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 [34]. Although CRFs can extract event structures, they aim to classify elements in a sequence (i.e., shots in a video). In other words, CRFs require training videos where each shot is annotated with an event's occurrence or absence, and cannot be used in the weakly supervised setting. Compared to this, HCRFs can handle this by performing CRF's shot labelling on hidden states, and combining labelling results to estimate the label of the entire video.

Finally, HCRFs have been successfully used in different applications such as object classification [21], action (gesture) recognition [32, 38], and audio analysis [11]. But, to the best of our knowledge, this paper describes the first application of HCRFs to MED. Because of this, in Section 4, we intensively investigate approaches for training effective HCRFs in terms of negative videos, parameter initialisation and hyperparameters (regularisation parameters).

3 Event Detection Method

This section presents our MED method using HCRFs. Its overview is firstly provided. Then, we explain a preprocessing method to recognise primitive meanings (concepts) in shots. An event is characterised by combining these meanings. Finally, we describe an event detection method based on HCRFs.

3.1 Overview

An event is 'highly-abstracted' in the sense that various objects interact with each other in different situa-

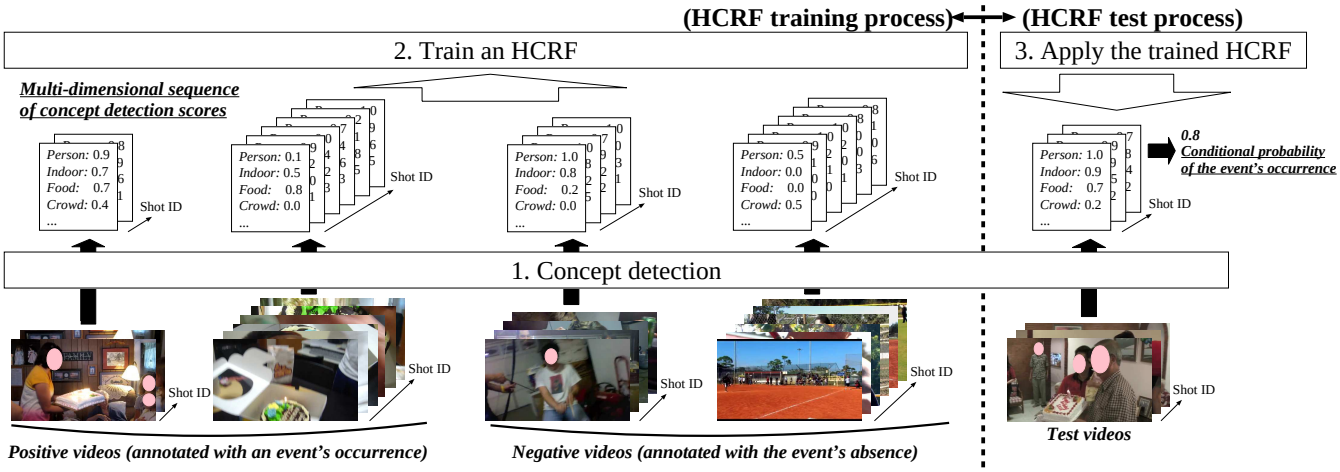


Fig. 2 An overview of our MED method where "birthday party" is used as an example event.

tions. In consequence, visual appearances of shots relevant to a certain event can be completely different. In other words, the set of these shots has got a huge variance in the space of low-level features like colour, edge, and motion. Hence, we adopt a *concept-based* approach which projects a shot into the space where each dimension represents the detection result of a concept [27]. Here, concepts are textual descriptions of meanings that can be observed from shots, such as objects like *Person* and *Car*, actions like *Walking* and *Airplane_Flying*, and scenes like *Beach* and *Nighttime*. In what follows, we denote concept names in italics to distinguish them from the other terms.

Owing to recent research progress, several concepts can be robustly detected irrespective of their sizes, directions and deformations in video frames. Thus, compared to the space of low-level features where each dimension just represents the physical value of a shot, in the space of concept detection results, each dimension represents the appearance of a human-perceivable meaning. In such a space, the variation of relevant shots to an event becomes smaller and can be modelled more easily. That is, relevant shots that are dissimilar at the level of low-level features, become more similar at the level of concepts.

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. In the bottom of Figure 2, each shot is represented by one video frame, and arranged from front to back based on its shot ID. Then, concept detection is conducted as a binary classification problem. For each concept, a detector is built using training shots, each annotated with its presence or absence. After that, the detector is used to associate every shot with a *detection score*, representing a scoring value between 0 and 1 in terms of the presence of the concept. A larger detection score indicates more likelihood that the concept is present in a shot.

Such detection scores are illustrated in the middle of Figure 2. For example, the first shot in the leftmost video shows an indoor scene where a person is bringing a birthday cake. Correspondingly, this shot is associated with the large detection scores 0.9, 0.7 and 0.7 for *Person*, *Indoor* and *Food*, respectively. Note that concept detection is uncertain because small (or large) detection scores for a concept may be falsely assigned to shots where it is actually present (or absent). Nonetheless, we assume that representative concepts in shots are successfully detected, and even if the detection of a concept fails on some shots, its contribution to an event can be appropriately evaluated by statistically analysing shots in positive and negative videos. For example, even though the shot exemplified above does not display *Crowd*, a relatively large detection score 0.4 is assigned to this shot. But, by checking the other shots in positive videos, it can be revealed that *Crowd* is irrelevant to the event "birthday party". Based on the above concept detection, we represent each video as a *multi-dimensional sequence* where each shot defined as a vector of detection scores is temporally ordered, as depicted in the middle of Figure 2.

Afterwards, an HCRF is trained using positive and negative videos, where an event's occurrence or absence is annotated only at the video level (i.e., weakly supervised setting). Hidden states are probabilistically optimised so as to maximise the discrimination between positive and negative videos. In an intuitive way, this optimisation can be thought as searching concepts and their relations which are typical for positive videos compared to negative ones. For example, in Figure 2, pos-

itive videos are more likely to contain shots with large detection scores for *Food* than negative videos. This leads a hidden state to favour the presence of *Food* with a high relevance value to the event "birthday party". In addition, due to the habit that people eat a birthday cake after candle blowing, positive videos often contain a shot sequence where a shot showing *Explosion_Fire* is followed by a shot showing *Food*. This associates a high relevance value with the transition between hidden states that favour the presence of *Explosion_Fire* and the one of *Food*, respectively. Like this, hidden states and state transitions for characterising relevant shots to the event are extracted under the weakly supervised setting. Meanwhile, event structures are characterised by hidden states and state transitions with high relevance values.

Finally, the trained HCRF is used to examine whether shots in each test video match with optimised hidden states and their transitions. As shown in the rightmost of Figure 2, the matching result is obtained as the conditional probability that the event occurs in the test video. The sorted list of test videos based on such conditional probabilities is returned as an MED result. Below, we describe the concept detection process and the HCRF training/test process.

3.2 Concept Detection

Our MED method characterises an event using appearances of concepts. Thus, the vocabulary of concepts should be sufficiently rich to describe 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 [20]. 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.

To build an accurate detector of each concept, we consider the following two issues: First, the concept appears in video frames with varied factors, such as its directions and deformations, and lighting conditions. A large number of training shots are required to cover such diverse appearances of the concept. Second, the concept does not necessarily appear in all video frames in a shot, and regions where it appears significantly vary depending on video frames. Hence, a feature needs to characterise various regions in many video frames. Since satisfying the listed issues requires expensive computational costs, we use the fast detector training/test method and the fast feature exaction method based on matrix operation [24]. The former realises batch computation of similarities among many training shots, and the latter allows batch computation of probability densities related to many regions in a shot. These methods make detector training/test and feature extraction about 10-37 and 5-7 times faster than the normal implementation, respectively.

Thanks to these accelerating methods, concept detection is conducted as follows (please refer to [24] for more detail): First, regions that are likely to characterise concept appearances, are sampled by applying Harris-Affine region detector to every other video frame in a shot. Subsequently, a Scale-Invariant Feature Transform (SIFT) descriptor is computed to quantify the appearance of each sampled region. Then, hundreds of thousands of SIFT descriptors obtained from the 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 detector using $30,000$ training shots. Here, training shots are collected from $545,872$ shots ($27,963$ videos) used at the TRECVID 2012 Semantic Indexing task [25], and the corresponding annotation data [4]. In total, detectors of 351 concepts are built because the annotation data contain more than one shot where the presence of each of these concepts is annotated.

3.3 Event Detection with HCRFs

Figure 3 illustrates an overview of an HCRF. It is important to note two different views of the HCRF. The first is the *model view* in Figure 3 (a) where the structure of the HCRF is shown, and the other is the *assignment view* in Figure 3 (b) where it is applied to a video. On first glance, the structure of the HCRF in Figure 3 (a) may appear similar to that of an HMM, however, because the HCRF is not a generative model but a discriminative model, hidden states represent characteristics of shots that are useful (or not-useful) for discriminating between positive and negative videos for an event. Such useful and not-useful shots are relevant and irrelevant shots to the event, respectively.

First, we define hidden states in an HCRF using the model view in Figure 3 (a). Let $y \in \{0,1\}$ be the event label where 0 and 1 mean an event's absence and occurrence, respectively. And, $\mathcal{H}$ denotes the set of all hidden states. In Figure 3 (a), the HCRF has four hidden states, so $\mathcal{H} = \{h_1, h_2, h_3, h_4\}$. As depicted in Figure 3 (a), each hidden state h_i ($1 \leq i \leq |\mathcal{H}|$) has the following three types of parameters: The first type is a 'label relevance' $\theta_{\text{label}}(y, h_i)$ representing the relevance of h_i to the event's absence ($y = 0$) or occurrence ($y = 1$).

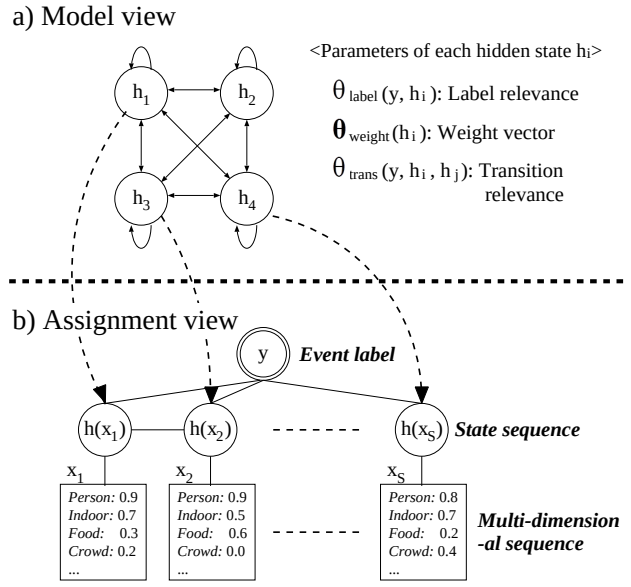


Fig. 3 An illustration of our HCRF model.

Hence, if h_i is assigned to a shot, $\theta_{\text{label}}(y, h_i)$ is used to represent the relevance (or irrelevance) of the shot to the event. The second type of parameter is a 'weight vector' $\boldsymbol{\theta}_{\text{weight}}(h_i)$ where each dimension indicates the weight of one concept. This vector represents characteristic concepts for h_i, and is used to match a shot with h_i in terms of concept appearances. The last type of parameter is a 'transition relevance' $\theta_{\text{trans}}(y, h_i, h_j)$ which represents the relevance of transition from h_i to h_j conditioned on the event label y. When h_i and h_j are assigned to two consecutive shots, $\theta_{\text{trans}}(y, h_i, h_j)$ examines whether this assignment is relevant in terms of the event's absence ($y = 0$) or occurrence ($y = 1$).

We explain how hidden states as defined above are assigned to shots in a video by referring to the assignment view in Figure 3 (b). Assuming 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 $\boldsymbol{x}_i$. Under this condition, $\boldsymbol{x}_i$ is assigned to a hidden state $h(\boldsymbol{x}_i) \in \mathcal{H}$. In Figure 3 (b), as depicted by dashed arrows from the model view, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_S$ are respectively assigned to h_1, h_3 and h_4, that is, $h(\boldsymbol{x}_1) = h_1$, $h(\boldsymbol{x}_2) = h_3$ and $h(\boldsymbol{x}_S) = h_4$. Then, the event label y is determined based on the sequence of hidden states $\boldsymbol{h}(\boldsymbol{x}) = (h(\boldsymbol{x}_1), \cdots, h(\boldsymbol{x}_S))^{\mathrm{T}}$ for $\boldsymbol{x}$.

Now let us evaluate the assignment of $\boldsymbol{h}(\boldsymbol{x})$ to $\boldsymbol{x}$ by assuming the event's occurrence ($y = 1$) or non-occurrence ($y = 0$) in $\boldsymbol{x}$. This is conducted by the following *potential function* $\Psi(y, \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}; \boldsymbol{\theta})$:

$$\Psi(y, \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{S} \theta_{\text{label}}(y, h(\boldsymbol{x}_i)) \tag{1}$$
$$+ \sum_{i=1}^{S} \boldsymbol{x}_i \cdot \boldsymbol{\theta}_{\text{weight}}(h(\boldsymbol{x}_i)) + \sum_{i=2}^{S} \theta_{\text{trans}}(y, h(\boldsymbol{x}_{i-1}), h(\boldsymbol{x}_i)),$$

where $\boldsymbol{\theta}$ is the whole set of parameters, consisting of label relevances $\theta_{\text{label}}(y, h_i)$s, weight vectors $\boldsymbol{\theta}_{\text{weight}}(h_i)$s and transition relevances $\theta_{\text{trans}}(y, h_i, h_j)$s for all hidden states. Equation (1) is based on the parameters of the hidden state $h(\boldsymbol{x}_i)$ assigned to the i-th shot $\boldsymbol{x}_i$. It is important to check the correspondence of a hidden state between the assignment and model views. For example, in Figure 3 (b), the first term of Equation (1) at $S = 1$ becomes $\theta_{\text{label}}(y, h_1)$ because $h(\boldsymbol{x}_1)$ in the assignment view corresponds to h_1 in the model view.

In Equation (1), $\Psi(y, \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}; \boldsymbol{\theta})$ combines the following three terms that evaluate $\boldsymbol{h}(\boldsymbol{x})$ from different perspectives: The first term sums label relevances of hidden states assigned to shots in $\boldsymbol{x}$. This represents the overall relevance of assigned hidden states to the event's occurrence ($y = 1$) or absence ($y = 0$). The second term accumulates the product between concept detection scores in $\boldsymbol{x}_i$ and the weight vector of its assigned hidden state. This term indicates the overall degree of how much shots in $\boldsymbol{x}$ match with assigned hidden states. The last term is the sum of transition relevances between hidden states assigned to two consecutive shots, and represents how relevant the transition of hidden states in $\boldsymbol{x}$ is to the event's occurrence or absence. By assuming that $\boldsymbol{h}(\boldsymbol{x})$ is appropriately selected for each training video, $\boldsymbol{\theta}$ should be optimised so that $\Psi(y = 1, \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}; \boldsymbol{\theta})$ is large for positive videos, while for negative ones $\Psi(y = 0, \boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}; \boldsymbol{\theta})$ is large.

Before implementing the optimisation process, we provide an intuitive explanation about how optimised hidden states contribute to managing the weakly supervised setting and extracting unclear event structures. Hidden states are optimised so as to discriminate between positive and negative videos. This leads $\boldsymbol{\theta}_{\text{weight}}(h_i)$s of some hidden states to characterise concepts that appear in several shots in positive videos, but hardly appear in shots in negative ones. Accordingly, these states are associated with large $\theta_{\text{label}}(y = 1, h_i)$s. In addition, concepts that hardly appear in shots in positive videos, are characterised by $\boldsymbol{\theta}_{\text{weight}}(h_i)$s of some hidden states with large $\theta_{\text{label}}(y = 0, h_i)$s. Thus, even in the weakly supervised setting, the HCRF can discriminate between relevant and irrelevant shots to an event using the above hidden states. Also, $\boldsymbol{\theta}_{\text{weight}}(h_i)$s of some hidden states represent concepts that appear in shots in both positive and negative videos. Such hidden states are assigned to

shots, for which neither relevance nor irrelevance can be determined. Meanwhile, event structures are captured by hidden states with high $\theta_{\text{label}}(y=1,h_i)$s and transitions with high $\theta_{\text{trans}}(y=1,h_i,h_j)$s.

The optimisation of hidden states is based on the following conditional probability of y given $\boldsymbol{x}$:

$$P(y|\boldsymbol{x},\boldsymbol{\theta}) = \sum_{\forall \boldsymbol{h}(\boldsymbol{x})\in\mathcal{H}} P(y,\boldsymbol{h}(\boldsymbol{x})|\boldsymbol{x},\boldsymbol{\theta}) \tag{2}$$

$$= \frac{\sum_{\forall \boldsymbol{h}(\boldsymbol{x})\in\mathcal{H}} e^{\Psi(y,\boldsymbol{h}(\boldsymbol{x}),\boldsymbol{x};\boldsymbol{\theta})}}{\sum_{\forall y'\in\mathcal{Y};\forall \boldsymbol{h}(\boldsymbol{x})\in\mathcal{H}} e^{\Psi(y',\boldsymbol{h}(\boldsymbol{x}),\boldsymbol{x};\boldsymbol{\theta})}}. \tag{3}$$

Equation (2) indicates that $\boldsymbol{h}(\boldsymbol{x})$ is marginalised out by taking the sum of $P(y,\boldsymbol{h}(\boldsymbol{x})|\boldsymbol{x},\boldsymbol{\theta})$s over all possible instances of $\boldsymbol{h}(\boldsymbol{x})$ (i.e., all possible assignments of hidden states to $\boldsymbol{x}$). The reason for this is that since hidden states cannot be observed, there is no direct assignment of shots to them. In other words, $\boldsymbol{x}_i$ does not necessarily match a single hidden state, but may match multiple states. Thus, the above mentioned marginalisation offers the 'soft-assignment' where $\boldsymbol{x}_i$ is assigned to every hidden state based on the probability of this assignment [35]. Equation (2) is further transformed into Equation (3), where the numerator with the fixed y is normalised by the denominator taking the sum of numerators with all $y' \in \{0,1\}$. Thus, Equation (3) can be considered as a conditional probability.

Regarding the computation of $P(y|\boldsymbol{x},\boldsymbol{\theta})$, the numerator and denominator in Equation (3) can be efficiently computed by the 'brief propagation' algorithm [21]. Here, for sequentially assigned hidden states like the ones in Figure 3 (b), all possible assignments of hidden states to $\boldsymbol{x}$ can be achieved by recursively accumulating all possible transitions of hidden states between two consecutive shots.

Finally, the optimisation of hidden states is conducted as follows: 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 $y^{(j)}$:

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{N} \log P(y^{(j)}|\boldsymbol{x}^{(j)},\boldsymbol{\theta}) - \frac{||\boldsymbol{\theta}||^2}{2\sigma^2} \quad , \tag{4}$$

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 (4) in terms of each parameter in $\boldsymbol{\theta}$ [21]. 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 (2). 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 on video data provided in TRECVID 2013 MED task [25, 28]. We used 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). For each event, an HCRF is trained using positive and negative videos collected from *EV* and *BG*, and tested on videos in *TE*. For reasons of simplicity, we call test videos containing the event *correct videos*. The MED result where test videos are ranked based on their relevance scores to the event, is evaluated by an Average Precision (AP). It is the average of precisions each of which is computed by cutting off the ranking of test videos at the position of a correct video. A larger AP means a better result where correct videos are ranked at higher positions. In addition, we use the Mean of APs (MAP) over all events as an overall evaluation measure.

Furthermore, we conduct the 'randomisation test' to check whether the performance difference between two methods is statistically significant or not [26]. For each event, APs of two methods are randomly swapped by assuming that there is no significant performance difference (null hypothesis). This produces a large number of pairs of MAPs for two methods. By referring to MAP differences for these pairs, it is examined whether the actual MAP difference is statistically unlikely or not. The threshold (significance level) for deciding the unlikelihood is set to 5% based on the usage of the randomisation test in TRECVID [25].

Table 1 summarises 10 events addressed in our experiments. The ID and description of each event are shown in the first and second columns, respectively. The third column indicates the average number of shots in positive videos. Note that all shots used to compute this average are not relevant to the event. Recall the weakly supervised setting where training videos are only labelled as whether an event is contained or not, and it is unknown which shots are relevant or irrelevant to the event. One of our main purposes is to examine the performance of HCRFs under the weakly supervised setting. Another main purpose is to investigate whether

Table 1 Events addressed in our experiments.

ID	Event Description	Avr. # of shots	# of correct videos
E006	Birthday party	10.69	186
E007	Changing a vehicle tire	10.32	111
E008	Flash mob gathering	25.12	132
E009	Getting a vehicle unstuck	5.38	95
E010	Grooming an animal	5.10	87
E011	Making a sandwich	14.06	140
E012	Parade	9.34	234
E013	Parkour	20.06	104
E014	Repairing an appliance	10.72	78
E015	Working on a sewing project	9.51	81

structures of events listed in Table 1 can be successfully extracted under the weakly supervised setting.

The forth column in Table 1 represents the number of correct videos. Here, 10 events other than those in Table 1 are also available. However, since numbers of correct videos for these events are too small (less than 33), we recognised that they are insufficient to examine the generality of whether correct videos with diverse visual appearances can be identified or not. In addition, APs are unstable for such small numbers of correct videos, because small changes of positions where correct videos are ranked, considerably affect averaging precisions at these positions. Thus, we only use 10 events in Table 1.

4.1 Tuning HCRFs

We set up HCRFs that will be compared to other methods and closely investigated after this section. In particular, we study conditions to train effective HCRFs, that is, what kind of negative videos should be used, and how to initialise parameters to be optimised.

4.1.1 Negative videos

According to the official instruction of the TRECVID MED task, for each event, 100 positive videos are collected from *EV*. Also, videos in *BG* have been proven to not-contain the event, so they are used as negative. Moreover, *EV* includes 'near-miss' videos which are visually similar to positive videos, but do not contain the event. Figure 4 depicts two examples of near-miss videos for the event "birthday party". *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. The reason is that many correct videos may be missed, because they may be similar to near-miss videos, and regarded as not-containing the event.

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

To investigate the effect of near-miss videos, two variants of HCRFs, *HCRF* and $HCRF_{\text{no-near}}$, are compared. For each event, *HCRF* constructs an HCRF using negative videos consisting of videos in *BG* and near-miss videos, while $HCRF_{\text{no-near}}$ only uses videos in *BG* as negative. Except for this, the condition is the same between *HCRF* and $HCRF_{\text{no-near}}$. Specifically, 100 positive videos are collected from *EV*. In addition, 10 hidden states are used, and due to the computational cost the maximum number of iterations for estimating $\boldsymbol{\theta}^*$ is set to 50. Initial parameters for $\boldsymbol{\theta}^*$ are determined by the parameter initialisation method described in the next section. It should be noted that the performance of an HCRF significantly depends on the parameter σ for L2 regularisation. Regarding this, we test each of $\sigma \in \{0.5, 1, 2, 4\}$ and select the one achieving the best performance. This aims to avoid under-estimating the performance of the HCRF. One solution for the σ selection will be presented in Section 4.4.

Table 2 shows the performance comparison between *HCRF* and $HCRF_{\text{no-near}}$. Each row presents APs for 10 events and the MAP over them. As can be seen from Table 2, the performance of *HCRF* is better than that of $HCRF_{\text{no-near}}$. However, this statement is only valid at the significance level of 10% in the randomisation test. Nonetheless, the performance improvement using near-miss videos for *E007*, *E008*, *E009*, *E012* and *E014*, seems much larger than the degradation without using near-miss videos for the other events. In other words, the advantage of using near-miss videos as negative seems to suppress its disadvantage. Thus, we use near-miss videos in the following experiments.

It should be noted that the above HCRFs have 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 [12]. Suppose that the number of negative videos (majority class) is much higher than that of positive ones (minority class). In this case, an HCRF which classifies almost all videos as negative may be constructed, because it can very accurately classify training videos containing the large number of negative ones.

We elaborate the effect of the imbalanced problem by comparing *HCRF* described above to $HCRF_{\text{sub}}$ built using a subset of negative videos, especially, randomly

Table 2 Performance comparison between *HCRF* and $HCRF_{\text{no-near}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	MAP
HCRF	0.0622	0.063	0.2317	0.1282	0.0369	0.0352	0.1062	0.1741	0.1797	0.0262	0.1043
$HCRF_{\text{no-near}}$	0.0717	0.0428	0.1954	0.1019	0.0476	0.0385	0.0929	0.1918	0.111	0.0224	0.0916

Table 3 Performance comparison between *HCRF* and $HCRF_{\text{sub}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	MAP
HCRF	0.0622	0.063	0.2317	0.1282	0.0369	0.0352	0.1062	0.1741	0.1797	0.0262	0.1043
$HCRF_{\text{sub}}$	0.0788	0.0627	0.1845	0.0930	0.0422	0.0544	0.1354	0.1502	0.0919	0.0273	0.0920
(Std. dev.)	±0.0058	±0.0092	±0.0115	±0.0151	±0.0069	±0.0075	±0.0147	±0.0124	±0.0195	±0.0080	
(Max)	0.0907	0.0754	0.2017	0.1191	0.0538	0.0709	0.1534	0.1790	0.1328	0.0467	
(Min)	0.0699	0.0484	0.1670	0.0649	0.0320	0.0435	0.0975	0.1353	0.0686	0.0194	

sampled 1,000 negative videos. Table 3 shows the performance comparison between *HCRF* and $HCRF_{\text{sub}}$. Considering the randomness of negative videos, for each event, $HCRF_{\text{sub}}$ is run 10 times using different sets of negative videos. The row denoted by $HCRF_{\text{sub}}$ presents the average of APs over 10 runs, and the MAP based on such averages for 10 events. The last three rows under this represent the standard deviation, maximum and minimum of APs in 10 runs.

The randomisation test states no significant performance difference between *HCRF* and $HCRF_{\text{sub}}$. This means that the imbalanced problem has no strong influence on the performance of HCRFs. In addition, as can be seen from Table 3, APs of $HCRF_{\text{sub}}$ considerably vary depending on negative videos. To avoid this variance and explicitly evaluate the effectiveness of HCRFs, we conduct the following experiments using all negative videos. Also, in Section 4.4, we will discuss how to utilise varied results with different negative videos for the performance improvement.

4.1.2 Parameter initialisation

Since the objective function in Equation (4) 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 [35]. First, an initial $\boldsymbol{\theta}$ is determined based on the 'hard-assignment' of hidden states to shots. 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 (2). Our method for $\boldsymbol{\theta}$ initialisation is summarised below.

First, for simplicity, each type of parameter constituting $\boldsymbol{\theta}$ is symbolised as follows (please refer to the model view in Figure 3 (a)): The set of label relevances of all hidden states is represented by $\boldsymbol{\theta}_{\text{label}} = \{\theta_{\text{label}}(y = 0, h_1), \cdots, \theta_{\text{label}}(y = 1, h_{|\mathcal{H}|})\}$, the set of weight vectors of all hidden states is denoted by $\boldsymbol{\theta}_{\text{weight}} = \{\boldsymbol{\theta}_{\text{weight}}(h_1), \cdots, \boldsymbol{\theta}_{\text{weight}}(h_{|\mathcal{H}|})\}$, and the one of transition relevances is described as $\boldsymbol{\theta}_{\text{trans}} = \{\theta_{\text{trans}}(y = 0, h_1, h_1), \cdots, \theta_{\text{trans}}(y = 1, h_{|\mathcal{H}|}, h_{|\mathcal{H}|})\}$. Since hidden states are shared by all shots, it is reasonable to initialise $\boldsymbol{\theta}_{\text{weight}}$ so as to characterise their distribution. Thus, 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 [9] is used. The weight vector of the i-th hidden state $\boldsymbol{\theta}_{\text{weight}}(h_i)$ is initialised using shots in the i-th cluster.

To initialise $\boldsymbol{\theta}_{\text{weight}}$ in a similar way to an HCRF [1], we construct a CRF that is a probabilistic model to predict the label of each element in a sequence [15]. For convenience, we call such labels *elemental labels*. The structure of the CRF is equivalent to that of the HCRF without the event label layer. In other words, the HCRF is an extension of the CRF where elemental labels are made hidden states to predict the label of the entire sequence. Thus, parameters characterising elemental labels in the CRF correspond to $\boldsymbol{\theta}_{\text{weight}}$ in the HCRF. Hence, we optimise the CRF by regarding the cluster index of each shot as its elemental label, and use optimised parameters characterising elemental labels as initial $\boldsymbol{\theta}_{\text{weight}}$. Note that although the optimisation of the CRF is similar to that of the HCRF, optimised parameters in the CRF are guaranteed as global optimum [15]. Thus, unlike the HCRF, we do not have to care initial parameters of the CRF.

In addition, using the optimised CRF, we compute the maximum likelihood sequence of elemental labels for each training video, and regard it as an assignment of hidden states. By only considering such an assignment for every training video, we initialise $\boldsymbol{\theta}_{\text{label}}$

[1] It is not reasonable to initialise $\boldsymbol{\theta}_{\text{weight}}(h_i)$ as the centre of the i-th cluster because of the difference of value ranges. While $\boldsymbol{\theta}_{\text{weight}}(h_i)$ take both positive and negative values, the cluster centre does not take negative ones because concept detection scores lie between 0 and 1.

Table 4 Performance comparison between *HCRF* and $HCRF_{\text{rand}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	MAP
HCRF	0.0622	0.063	0.2317	0.1282	0.0369	0.0352	0.1062	0.1741	0.1797	0.0262	0.1043
$HCRF_{\text{rand}}$	0.0606	0.0403	0.2043	0.0710	0.0377	0.0434	0.1058	0.1822	0.1346	0.0249	0.0905

and $\boldsymbol{\theta}_{\text{trans}}$ in the HCRF framework. Here, $P(y|\boldsymbol{x},\boldsymbol{\theta})$ in Equation (2) and (3) can be simplified as follows [38]:

$$P(y|\boldsymbol{x},\boldsymbol{\theta}) = \frac{e^{\Psi(y,\boldsymbol{h}_{\text{crf}}(\boldsymbol{x}),\boldsymbol{x};\boldsymbol{\theta})}}{\sum_{\forall y' \in \mathcal{Y}} e^{\Psi(y',\boldsymbol{h}_{\text{crf}}(\boldsymbol{x}),\boldsymbol{x};\boldsymbol{\theta})}}, \quad (5)$$

where for a training video $\boldsymbol{x}$, $\boldsymbol{h}_{\text{crf}}(\boldsymbol{x})$ is the assignment of hidden states obtained based on the optimised CRF. According to this, the objective function in Equation (4) can be decomposed as follows:

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{N} \Psi(y^{(j)}, \boldsymbol{h}_{\text{crf}}^{(j)}(\boldsymbol{x}^{(j)}), \boldsymbol{x}^{(j)}; \boldsymbol{\theta}) - \sum_{j=1}^{N} \log \sum_{\forall y' \in \mathcal{Y}} e^{\Psi(y', \boldsymbol{h}_{\text{crf}}^{(j)}(\boldsymbol{x}^{(j)}), \boldsymbol{x}^{(j)}; \boldsymbol{\theta})} - \frac{||\boldsymbol{\theta}||^2}{2\sigma^2}. \quad (6)$$

As proven in [38], this objective function is convex. Initial $\boldsymbol{\theta}_{\text{label}}$ and $\boldsymbol{\theta}_{\text{trans}}$ that are global optimum can be obtained by a conventional gradient ascend method. Finally, $\boldsymbol{\theta}$ consisting of $\boldsymbol{\theta}_{\text{weight}}$, $\boldsymbol{\theta}_{\text{label}}$ and $\boldsymbol{\theta}_{\text{trans}}$ initialised above, is refined to $\boldsymbol{\theta}^*$ by the original HCRF optimisation in Equation (4).

We examine the effectiveness of this initialisation method by comparing *HCRF* which uses this method, to $HCRF_{\text{rand}}$ where $\boldsymbol{\theta}$ is initialised with random values [21]. Table 4 presents their performance comparison. Considering the variance in the performance of $HCRF_{\text{rand}}$, for each event, we run it 10 times using $\boldsymbol{\theta}$ initialised by different random values. The average APs of these 10 runs is shown in the row for $HCRF_{\text{rand}}$. Table 4 demonstrates that *HCRF* outperforms $HCRF_{\text{rand}}$, which is statistically confirmed with the significance level of 5%. In addition, on average, the objective function value using $\boldsymbol{\theta}^*$ optimised by *HCRF* is 8.51% larger than the one using $\boldsymbol{\theta}^*$ optimised by $HCRF_{\text{rand}}$. That is, the former is better founded than the latter from the computational perspective. Therefore, the above initialisation is useful for building effective and valid HCRFs.

4.2 Evaluation for the Weakly Supervised Setting

To examine the effectiveness of HCRFs under the weakly supervised setting, we compare *HCRF* established in the previous section to the following three methods:

1. SVM_{avr}: As seen from Equation (1), 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. The SVM parameter for penalising mis-classified training videos has been heuristically set to 2. Using SVM_{avr}, we aim to examine the effectiveness of *HCRF* where hidden states are used to precisely characterise events based on shot-level vectors.

2. SVM_{max}: Similar to SVM_{avr}, this constructs a linear SVM based on video-level vectors by max-pooling.

3. *HMM*: HMMs are the most popular model to classify sequential data based on their structures, and have several common characteristics to HCRFs. But, for each event, an HMM can model structures only in positive (or negative) videos where all shots are regarded as relevant (or irrelevant). Hence, by comparing *HMM* to *HCRF*, we aim to examine whether the latter can appropriately discriminate between relevant and irrelevant shots to an event.

We construct *HMM* with the following configuration using Hidden Markov Model Toolkit (HTK) [35]: Like *HCRF*, *HMM* permits the transition between any pair of hidden states (i.e., ergodic HMM). In addition, as each hidden state in *HCRF* matches shots using one function (i.e., $\boldsymbol{x}_i \cdot \boldsymbol{\theta}_{\text{weight}}(h_i)$), each state in *HMM* uses a single mixture (normal distribution). Due to the quadratic increase of parameters, each concept (dimension) is assumed to be independent of each other. Moreover, since a normal distribution involves the variance in its denominator, concepts where the variance of detection scores is too small (less than 10^{-6}) are removed. As a result, each shot is represented as a vector of detection scores for 343 concepts. In the training process, *HMM* prohibits any variance of each normal distribution to be less than 0.01. In the test process, using the log-space forward algorithm [19], *HMM* computes the log-probability that a test video is generated by the trained HMM. That is, hidden states are marginalised out like Equation (2).

Furthermore, *HMM* detects each event using two types of HMMs, HMM_{pos} and HMM_{neg}, each type is built on positive or negative videos (please refer to Section 2 for the rationale of HMM_{neg}). The relevance score of a test video to the event is determined by subtracting the log-probability by HMM_{neg} from the one by HMM_{pos}. In particular, the best number of hidden states is unknown for HMM_{pos} and HMM_{neg}. Thus, we build 15 variants of HMM_{pos} and those of HMM_{neg},

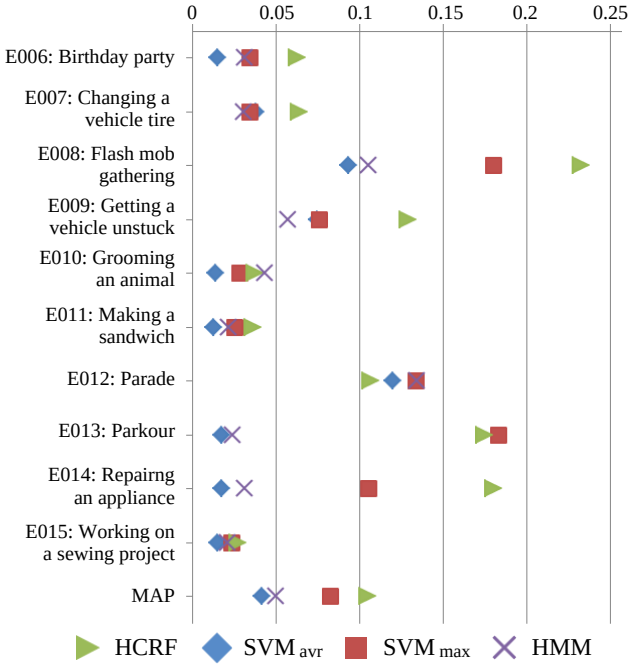


Fig. 5 Performance comparison between $HCRF$, SVM_{avr}, SVM_{max} and HMM.

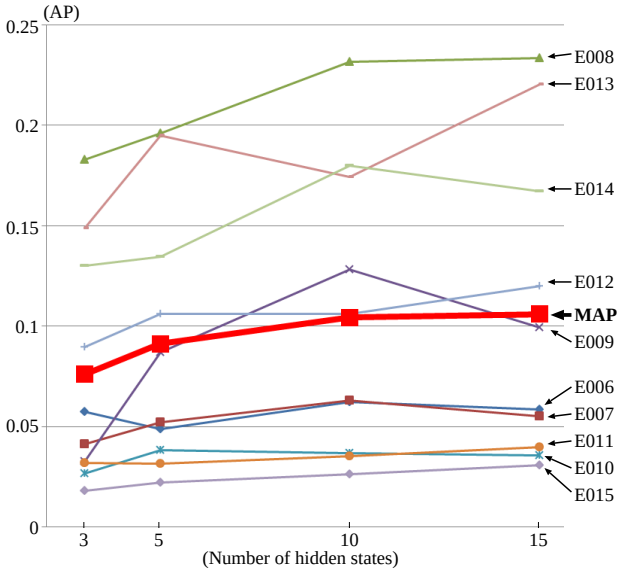


Fig. 6 Performance transition using different numbers of hidden states.

corresponding to using 1 to 15 hidden states. Then, by testing all possible pairs of them, we select the best one. This can be assumed as the upper-bound performance of HMM. According to our preliminary experiment, the above configuration has been confirmed as the best [2].

Figure 5 shows the performance comparison between $HCRF$, SVM_{avr}, SVM_{max}, and HMM. 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 MAPs over all 10 events. As can be seen from Figure 5, for 7 of 10 events, $HCRF$ outperforms the other three methods. The randomisation test has confirmed that $HCRF$ is superior to SVM_{avr}, SVM_{max} and HMM with the significance level of 3%. The superiority of $HCRF$ over SVM_{avr} and SVM_{max} validates the effectiveness of the precise shot-level characterisation of events, and its superiority over HMM verifies that relevant and irrelevant shots to an event are appropriately distinguished. Therefore, HCRFs are very effective for the weakly supervised setting.

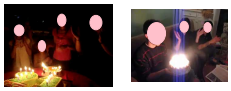
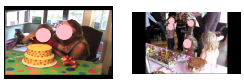
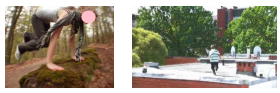
[2] We also tested PCA to make each dimension (concept) independent of each other, and the normalisation to obtain uniformed dimensions with the mean zero and the variance one. However, neither of them worked well. It can be considered that detection scores for each concept are appropriately biased by the detector, so editing their distribution does not offer improvement.

4.3 Evaluation for Extracting Unclear Event Structures

We investigate the usefulness of HCRFs for extracting 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 larger diversity of shots relevant (or irrelevant) to an event. Figure 6 shows the transition of performances depending on different numbers of hidden states. As shown in the horizontal axis, APs and MAPs obtained by 3, 5, 10 and 15 hidden states are plotted. Here, except for the number of hidden states, each AP is obtained by the same configuration to that of $HCRF$ in Section 4.1. Figure 6 indicates that although the performance improvement is relatively unclear at the event level, the overall performance (i.e., MAPs plotted by the bold line) is gradually improved using a larger number of hidden states. This implies that a larger diversity of shots can be covered using more hidden states. However, as exposed by the randomisation test, the performance using 15 hidden states is not statistically significant compared to the one using 10 states. Thus, as the trade-off between the performance and the computational cost, using 10 hidden states is considered reasonable.

Next, we check event structures extracted by HCRFs in terms of shots characterised by hidden states. Table 5 shows two hidden states which characterise shots relevant to *E006*, *E009* or *E013*, as depicted by column names. They are a part of 10 hidden states extracted by $HCRF$ in Figure 6. Each hidden state is represented

Table 5 An Illustration of extracted hidden states which characterise relevant shots to *E006*, *E009* and *E013*.

	E006: Birthday party	*E009:Getting a vehicle unstuck*	*E013:Parkour*
$\theta_{\text{label}}(y, h_i)$	$y = 0{:}-0.396$, $y = 1{:}0.228$	$y = 0{:}-0.548$, $y = 1{:}0.293$	$y = 0{:}-0.101$, $y = 1{:}0.571$
$\boldsymbol{\theta}_{\text{weight}}(h_i)$	0.247 (*Moonlight*) 0.204 (*Nighttime*) 0.192 (*Entertainment*) 0.125 (*Event*) 0.121 (*Singing*) 0.097 (*Celebrity_Entertainment*) 0.093 (*Dancing*) 0.093 (*Instrumental_Musician*) 0.057 (*Person*) 0.056 (*Face*)	1.665 (*Text_On_Artificial_Background*) 1.421 (*Waterscape_Waterfront*) 1.342 (*Head_And_Shoulder*) 1.316 (*Car*) 1.208 (*Infants*) 1.112 (*Outdoor*) 1.085 (*Adult_Male_Human*) 1.081 (*Daytime_Outdoor*) 1.065 (*Driver*) 1.051 (*Human_Young_Adult*)	1.634 (*Indoor*) 1.236 (*Building*) 0.886 (*Overlaid_Text*) 0.844 (*Bridges*) 0.768 (*Graphic*) 0.727 (*Door_Opening*) 0.726 (*Windows*) 0.654 (*Animation_Cartoon*) 0.631 (*Legs*) 0.589 (*Sky*)
Example shots	 (Candle blowing)	 (Stuck in water)	 (Jumping over buildings)
$\theta_{\text{label}}(y, h_i)$	$y = 0{:}-1.169$, $y = 1{:}0.470$	$y = 0{:}-2.645$, $y = 1{:}0.222$	$y = 0{:}-0.196$, $y = 1{:}0.005$
$\boldsymbol{\theta}_{\text{weight}}(h_i)$	1.157 (*Outdoor*) 1.073 (*Sofa*) 1.047 (*Room*) 1.029 (*Boy*) 1.011 (*Female_Person*) 0.997 (*Two_People*) 0.968 (*Dining_Room*) 0.941 (*Girl*) 0.845 (*Singing*) 0.782 (*Food*)	2.412 (*Ground_Vehicles*) 2.157 (*Vertebrate*) 2.093 (*Road*) 1.703 (*Civilian_Person*) 1.620 (*Van*) 1.597 (*Face*) 1.542 (*Human_Young_Adult*) 1.517 (*Swimming_Pools*) 1.490 (*Car*) 1.395 (*Singing*)	0.361 (*Forest*) 0.292 (*Explosion_Fire*) 0.243 (*Urban_Park*) 0.239 (*Trees*) 0.207 (*Plant*) 0.200 (*Sunny*) 0.184 (*Vegetation*) 0.154 (*Building*) 0.152 (*Suburban*) 0.150 (*Throwing*)
Example shots	 (Chat in parties)	 (Vans are often stuck)	 (Parkour in forests/suburbs)

by a set of rows denoted by $\theta_{\text{label}}(y, h_i)$, $\boldsymbol{\theta}_{\text{weight}}(h_i)$ and "Example shots". As can be seen from rows of $\theta_{\text{label}}(y, h_i)$, every hidden state has a higher label relevance to an event's occurrence than the relevance to its non-occurrence (i.e., $\theta_{\text{label}}(y = 0, h_i) < \theta_{\text{label}}(y = 1, h_i)$). A row named as $\boldsymbol{\theta}_{\text{weight}}$ show 10 concepts with the highest weights, as represented by numbers on the left of concept names. Examples of shots characterised by such concepts are shown in rows of "Example shots".

As can be seen from Table 5, hidden states appropriately represent characteristic shots for each event's occurrence. For example, upper hidden states for *E006*, *E009* and *E013* can be considered to characterise shots showing candle blowing scenes, scenes where cars are stuck in water, and scenes where persons jump over buildings, respectively. This means that HCRFs can extract descriptions of characteristic shots for an event, under the weakly supervised setting where videos are only annotated as whether the event is contained or not (no shots are annotated). Thus, using HCRFs, we can create a knowledge base storing such descriptions for various events with reduced annotation effort. In addition, extracting human-perceivable event descriptions under the weakly supervised setting, is difficult for existing MED methods such as linear SVMs on low-level features [1, 18], non-linear SVMs [7, 13], SVMs with latent variables [16, 31], and Fisher kernel encoding [29] (see Section 2).

Now, we explore the temporal structures of events by comparing *HCRF* to $HCRF_{\text{no}}$. One may think that since there are no clear temporal structures of events in unconstrained videos created by arbitrary editing techniques, it is meaningless to consider the relation between two consecutive shots based on transitions among hidden states. Thus, $HCRF_{\text{no}}$ does not consider state transitions by removing the term $\sum \theta_{\text{trans}}(y, h(\boldsymbol{x}_{i-1}), h(\boldsymbol{x}_i))$ from Equation (1). Table 6 shows the performance comparison between *HCRF* and $HCRF_{\text{no}}$, where the former significantly outperforms the latter.

In particular, $HCRF_{\text{no}}$ frequently causes false positive detection. Many videos where an event does not occur are falsely detected, only because they contain shots that 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 for children). For *E009: Getting a vehicle unstuck*, shots in falsely detected videos just show cars. Compared to this, *HCRF*

Table 6 Performance comparison between $HCRF$ and $HCRF_{\mathrm{no}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	MAP
$HCRF$	0.0622	0.063	0.2317	0.1282	0.0369	0.0352	0.1062	0.1741	0.1797	0.0262	0.1043
$HCRF_{\mathrm{no}}$	0.0305	0.0406	0.0777	0.0612	0.0305	0.0211	0.1228	0.0268	0.0224	0.0133	0.0446

Table 7 Performance comparison among $HCRF$, $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$.

	E006	*E007*	*E008*	*E009*	*E010*	*E011*	*E012*	*E013*	*E014*	*E015*	MAP
$HCRF$	0.0622	0.063	0.2317	0.1282	0.0369	0.0352	0.1062	0.1741	0.1797	0.0262	0.1043
$HCRF^{(\sigma)}_{\mathrm{bag}}$	**0.0651**	0.0509	**0.2625**	0.0726	0.032	**0.0415**	0.0984	**0.2512**	**0.2064**	0.0236	0.1104
$HCRF^{(\sigma,n)}_{\mathrm{bag}}$	**0.0969**	**0.0832**	0.2006	**0.1372**	0.0318	**0.0601**	**0.1604**	**0.2202**	0.1625	**0.0293**	0.1182

indicates that for *E006*, shots where children appear are often followed by shots containing *Singing* or *Dancing*. In addition, for *E009*, shots displaying *Car* and *Road* are not repeated, but tend to be followed by shots showing *Snow* which causes stuck of *Car*. Thus, transitions among hidden states are effective constraints to alleviate false positive detection. In other words, temporal structures of events exist even in unconstrained videos, and can be captured by transitions among hidden states. In Section 5, we will discuss how to extract the longer relation of shots than the one between two consecutive shots.

4.4 Bagging of HCRFs

In Section 4.1.1, we described two factors, σ and a set of negative videos, which cause unstable results of HCRFs. We acquired one finding on these results, where correct videos are ranked at relatively high positions, while incorrect ones 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 data [6], we combine results by different σs and different sets of negative videos into a single result.

We have devised two bagging approaches, $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$. For an event, $HCRF^{(\sigma)}_{\mathrm{bag}}$ 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 $HCRF^{(\sigma,n)}_{\mathrm{bag}}$, 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 $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$, for each test video $\boldsymbol{x}$, the sum of $P(y = 1|\boldsymbol{x}, \boldsymbol{\theta}^*)$s obtained by different HCRFs, is simply used as the final relevance score to the event.

Table 7 shows the performance comparison between the bagging approaches, and $HCRF$ in Section 4.1 where the best σ is manually selected. In Table 7, APs in bold font indicate that $HCRF^{(\sigma)}_{\mathrm{bag}}$ or $HCRF^{(\sigma,n)}_{\mathrm{bag}}$ outperforms $HCRF$. Overall, as seen from the column *MAP*, both of $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$ are more accurate than $HCRF$. Although no significant performance difference among them is indicated by the randomisation test ($HCRF^{(\sigma,n)}_{\mathrm{bag}}$ is superior to $HCRF$ with the significance level of 8%), the important finding is that bagging leads to similar or even superior results to the ones obtained by manual.

Finally, Table 7 presents different characteristics of $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$. Except *E009*, $HCRF^{(\sigma)}_{\mathrm{bag}}$ yields great improvement on *E008*, *E013* and *E014*, while its performance is similar to that of $HCRF$ on the other events. Thus, bagging with different σs and all negative videos, works quite well on some events. On the other hand, $HCRF^{(\sigma,n)}_{\mathrm{bag}}$ offers modest improvement on most events, but it is significantly degraded on some events like *E008* and *E014*, due to insufficient negative videos to build each HCRF. Hence, $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$ can be considered as complementary. One interesting research topic is how to select the best bagging strategy depending on events. If the best of $HCRF^{(\sigma)}_{\mathrm{bag}}$ and $HCRF^{(\sigma,n)}_{\mathrm{bag}}$ could be correctly selected for each event in Table 7, the MAP would become 0.131.

5 Conclusion and Future Work

In this paper, we addressed the weakly supervised setting and unclear event structures in MED, and introduced a method using HCRFs. In an HCRF, hidden states are used as the intermediate layer to examine the relevance of each shot to an event. In addition, each hidden state represents characteristic concepts and has the relation to the other states. Such hidden states are probabilistically optimised so as to discriminate between positive and negative videos for the event. As a result, even in the weakly supervised setting, hidden states appropriately distinguish relevant shots from irrelevant ones. Moreover, the event structure is captured

by concepts specific to hidden states and their relation. In the experiments, we have showed adequate tuning of HCRFs, their effectiveness for the weakly supervised setting and unclear event structures, and the improvement using the bagging approach.

In the future, we will explore the following three issues: First, in Table 5, while large weights are successfully assigned to concepts which are related to events, they are also assigned to several not-related concepts, such as *Infants* in the upper hidden state for *E009*, and *Animation_Cartoon* in the upper hidden state for *E013*. One main reason is the current imperfect concept detection using only a single image feature (SIFT). Thus, we will incorporate motion and audio features into concept detection to improve its performance [24]. This will also yield the improvement of the MED performance.

Second, based on the flexibility of the potential function in Equation (1), HCRFs can deal with long-range dependencies among shots using an 'window feature' [32]. This represents each shot as the concatenation of concept detection scores in the previous, current and next shots. Thereby, the relation between two consecutive shots can count shots that are separated by more than one shot. Although we tested the window feature, the MAP over the 10 events in Section 4 was 0.1000, which has no significant difference to the MAP 0.1043 of *HCRF* based on dependencies between two consecutive shots. One main reason is that the temporal order of shots is often corrupted by inserting shots, that display different meanings than those of surrounding shots. Hence, in order to incorporate long-range dependencies among shots into HCRFs, we need to flexibly treat the distorted order of shots.

To this end, we will examine the following two methods. The first one models the temporal continuity of a concept's presence based on time series segmentation [23]. A video is divided into shot sequences each of which is characterised by a probabilistically distinct pattern of the concept's presence. As a result, the low detection score for the concept in a shot is flexibly modified by considering the ones in surrounding shots. Then, HCRFs are built using the above modified detection scores. The second method is to represent a video as a tree, where shots that are visually similar and temporally close to each other, are gradually grouped into nodes [22]. A node in the tree is represented by applying the max-pooling (or average-pooling) to concept detection scores of shots in that node. After that, HCRFs are trained where, in addition to the sequential video representation used in this paper, the belief propagation algorithm works for the above tree-structured video representation to efficiently compute the conditional probability of an event's occurrence or absence (Equation (2)) [21].

Finally, the results in Section 4.3 (especially, Table 5) indicate the possibility that meaningful event structures can be extracted from unconstrained Web videos. Thus, the last long-term issue is to build a knowledge base about events. Various event structures extracted from a large amount of Web videos, may be compared and structured using a pattern recognition or data mining method.

Acknowledgements The research work by Kimiaki Shirahama leading to this article has been funded by the Postdoctoral Fellowship for Research Abroad by Japan Society for the Promotion of Science (JSPS). Also, this work was in part supported by JSPS through Grand-in-Aid for Scientific Research (B): KAKENHI (26280040).

References

1. Aly R *et al* (2012) AXES at TRECVid 2012: KIS, INS, and MED. In: Proc. of TRECVID 2012, URL http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/axes.pdf
2. Ando R, Shinoda K, Furui S, Mochizuki T (2006) Robust scene recognition using language models for scene contexts. In: Proc. of MIR 2006, pp 99–106
3. Arijon, D (1976) Grammar of the Film Language. Silman-James Press
4. Ayache S, Quénot G (2008) Video corpus annotation using active learning. In: Proc. of ECIR 2008, pp 187–198
5. Barnard M, Odobez J (2005) Sports event recognition using layered HMMs. In: Proc. of ICME 2005, pp 1150–1153
6. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
7. Cheng H *et al* (2012) SRI-Sarnoff AURORA system at TRECVID 2012: Multimedia event detection and recounting. In: Proc. of TRECVID 2012, URL http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/aurora.pdf
8. Davenport G, Smith TA, Pincever N (1991) Cinematic primitives for multimedia. IEEE Comput Graph Appl 11(4):67–74
9. Fujisawa M (2012) Bayon - A Simple and Fast Clustering Tool. URL http://code.google.com/p/bayon/
10. Gemmell DJ, Vin HM, Kandlur DD, Rangan PV, Rowe LA (1995) Multimedia storage servers: A tutorial. IEEE Comput 28(5):40–49
11. Gunawardana A, Mahajan M, Acero A, Platt JC (2005) Hidden conditional random fields for phone classification. In: Proc. of INTERSPEECH 2005, pp 1117–1120
12. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
13. Inoue N, Wada T, Kamishima Y, Shinoda K, Sato S (2011) TokyoTech+Canon at TRECVID 2011. In: Proc. of TRECVID 2011, URL http://www-nlpir.nist.gov/projects/tvpubs/tv11.papers/tokyotechcanon.pdf
14. Jiang YG, Bhattacharya S, Chang SF, Shah M (2013) High-level event recognition in unconstrained videos. Int J Multimed Inf Retr 2(2):73–101

15. Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of ICML 2001, pp 282–289
16. Li W, Yu Q, Divakaran A, Vasconcelos N (2013) Dynamic pooling for complex event recognition. In: Proc. of ICCV 2013, pp 2728–2735
17. Li X, Snoek CGM (2009) Visual categorization with negative examples for free. In: Proc. of MM 2009, pp 661–664
18. Liu J, McCloskey S, Liu Y (2012) Local expert forest of score fusion for video event classification. In: Proc. of ECCV 2012, pp 397–410
19. Mann TP (2006) Numerically Stable Hidden Markov Model Implementation. URL http://bozeman.genome.washington.edu/compbio/mbt599_2006/hmm_scaling_revised.pdf, HMM Scaling Tutorial
20. Naphade M *et al* (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91
21. Quattoni A, Wang S, Morency L, Collins M, Darrell T (2007) Hidden conditional random fields. IEEE Trans Pattern Anal Mach Intell 29(10):1848–1852
22. Rui Y, Huang TS, Mehrotra S (1999) Constructing table-of-content for videos. Multimed Syst 7(5):359–368
23. Shirahama K, Uehara K (2008) A novel topic extraction method based on bursts in video streams. Int J Hybrid Inf Technol 1(3):21–32
24. Shirahama K, Uehara K (2012) Kobe university and Muroran institute of technology at TRECVID 2012 semantic indexing task. In: Proc. of TRECVID 2012, URL http://www-nlpir.nist.gov/projects/tvpubs/tv12.papers/kobe-muroran.pdf
25. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proc. of MIR 2006, pp 321–330
26. Smucker MD, Allan J, Carterette B (2007) A comparison of statistical significance tests for information retrieval evaluation. In: Proc. of CIKM 2007, pp 623–632
27. Snoek CGM, Worring M (2009) Concept-based video retrieval. Found Trends Inf Retr 2(4):215–322
28. Strassel, S *et al* (2012) Creating HAVIC: Heterogeneous audio visual internet collection. In: Proc. of LREC 2012, pp 2573–2577
29. Sun C, Nevatia R (2013) ACTIVE: Activity concept transitions in video event classification. In: Proc. of ICCV 2013, pp 913–920
30. Tanaka K, Ariki Y, Uehara K (1999) Organization and retrieval of video data. IEICE Trans Inf Syst 82(1):34–44
31. Vahdat A, Cannons K, Mori G, Oh S, Kim I (2013) Compositional models for video event detection: A multiple kernel learning latent variable approach. In: Proc. of ICCV 2013, pp 1185–1192
32. Wang SB, Quattoni A, Morency L, Demirdjian D, Darrell T (2006) Hidden conditional random fields for gesture recognition. In: Proc. of CVPR 2006, pp 1521–1527
33. Wang T, Li J, Diao Q, Hu W, Zhang Y, Dulong C (2006) Semantic event detection using conditional random fields. In: Proc. of CVPRW 2006, p 109
34. Yin J, Hu DH, Yang Q (2009) Spatio-temporal event detection using dynamic conditional random fields. In: Proc. of IJCAI 2009, pp 1321–1326
35. Young S et al (2009) The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department, URL http://htk.eng.cam.ac.uk/
36. Yu H, Han J, Chang KC (2004) PEBL: Web page classification without negative examples. IEEE Trans Knowl Data Eng 16(1):70–81
37. Zhai Y, Rasheed Z, Shah M (2004) A framework for semantic classification of scenes using finite state machines. In: Proc. of CIVR 2004, pp 279–288
38. Zhang J, Gong S (2010) Action categorization with modified hidden conditional random field. Pattern Recognit 43(1):197 – 203