Abstract—The task of recognizing events from video has attracted a lot of attention in recent years. However, due to the complex nature of user-defined events, the use of purely audio-visual content analysis without domain knowledge has been found to be grossly inadequate. In this paper, we propose to construct a semantic-visual knowledge base to encode the rich event-centric concepts and their relationships from the well-established lexical databases, including FrameNet, as well as the concept-specific visual knowledge from ImageNet. Based on this semantic-visual knowledge bases, we design an effective system for video event recognition. Specifically, in order to narrow the semantic gap between the high-level complex events and low-level visual representations, we utilize the event-centric semantic concepts encoded in the knowledge base as the intermediate-level event representation, which offers both human-perceivable and machine-interpretable semantic clues for event recognition. In addition, in order to leverage the abundant ImageNet images, we propose a robust transfer learning model to learn the noise-resistant concept classifiers for videos. Extensive experiments on various real-world video datasets demonstrate the superiority of our proposed system as compared to the state-of-the-art approaches.

Index Terms—Concept detection, event recognition, knowledge base.

I. INTRODUCTION

Due to the explosive growth of user-generated videos shared on the Internet and the potential demands in various multimedia applications, such as the Web video indexing, consumer content management, and open-source intelligence analysis, the task of multimedia event recognition has received increasing attention in recent years. The multimedia event recognition involves the automatic recognition of complex events from a large set of unconstrained videos. This task is extremely challenging due to four factors: (a) multimedia events are higher-level complex descriptions of multimedia data, which include several semantic concepts of objects, scenes and human actions; (b) events often have significant intra-class variations and inter-class similarities on both the visual appearance and semantic concepts; for example, both the events of ‘birthday party’ and ‘wedding ceremony’ may contain objects such as candles and cakes as shown in Fig. 1, but these objects might not appear in all the videos that belong to the same event category; (c) videos are often taken by consumers in unconstrained environments with different recording devices; as a result they usually contain significant visual variations and noise [1]; and (d) there are typically only few positive examples for complicated events. Many efforts are made on evaluating the efficacy of low-level features [2], [3]. However, as events are often characterized by similarity in semantics rather than visual appearance, recent approaches start to use high-level semantic concepts to assist in the recognition of events [4]–[10]. Many of these recent approaches introduce the use of a bank of concepts representation for this task [8], [4], [10], [11].

Studies [12], [7], [4] have demonstrated that some concepts are more useful than others in recognizing certain events. However, the key question is which concepts are most useful in identifying certain events and how to identify the concepts within the video stream. Current domain-free approaches learn the event-discriminative concepts from video examples [7], [5], [13]. When a sufficient number of and strong labelled training samples is provided, these learning methods are able to deliver promising results. However, in many of the semantically rich events, there are often few positive examples accompanied with significant intra-class variations, while the negative examples come from an infinite semantic space. To tackle such problems, there is a need to formulate the semantic concept-event relationship in a generalized way based on...
human’s broad understanding of events. Some existing datasets [14] are accompanied by natural language descriptions of the complex events, where there are coarse high-level descriptions of the concepts occurring in the events; however they often lack visual clues for these concepts. Hence, we need to obtain the visual information about the concepts to facilitate the visual recognition task.

Fortunately, there are lots of readily available concept structures in the Natural Language Processing (NLP) community to tackle the problem of concept-event association. A very rich semantic resource named FrameNet [15] has been developed to provide information on the type of an event, its description, along with the participants in it. While the existing pictorially enriched ontology ImageNet [16], which includes visual concepts together with linguistic keywords from WordNet [17], has been developed to organize the semantic concepts from general to specific according to their relationships.

Learning the potentially infinite set of semantic concepts from multimedia data is difficult, which in turn makes the link between events and the potential visual concepts a challenging task. Recent work in computational linguistics has shown that it is possible to extract semantic vectors representing concepts from images [18]. However, due to the deficiency of labelled concepts on video key frames, most of the existing concept detectors used in video event detection task [13], [8], [10] still utilize user-tagged Web images. In summary, learning the concepts from the auxiliary Web images for video task present two major problems: (a) the auxiliary images and target videos are usually of different qualities and from different domains, which will lead to great difference between their feature distributions; and (b) there are a lot of noise in concept assignment in the available Web images [19].

Motivated by the above observations and the available knowledge resources, we propose a novel video event recognition framework as shown in Fig. 2. The use of semantic relationship between concepts and events will enable us to incorporate more informative prior knowledge into video event recognition. Hence we construct a semantic-visual knowledge base, which encodes a rich set of event-centric concept relationship. In particular, we employ the well-established lexical knowledge base named FrameNet to align video events to the nodes (known as Frames) in FrameNet, and concepts in its constituent Lexical Units (LUs). We further extend the coverage of LUs in FrameNet to model visual concepts by using ImageNet according to the context of the LUs. Based on this proposed semantic-visual knowledge base, we devise an effective video event recognition system, which employs the event-centric concepts as intermediate presentation to narrow the semantic gap. In order to overcome the problem of lack of training samples for concept learning in videos, we propose a robust transfer learning model by leveraging the abundant labelled ImageNet images. We devise a multiple kernel learning algorithm to minimize the domain difference between video and image spaces.

The contributions of this paper are as follows.

- We automatically construct a semantic-visual knowledge base that contains rich event-centric concept relationships. The knowledge base is derived from the well-established lexical database named FrameNet as well as the rich concept-specific visual knowledge from ImageNet.
- In order to effectively model the event-centric semantic concepts to facilitate the video event recognition, we propose a robust transfer learning model based on stochastic programming. The proposed model learns the noise-resistant concept classifiers using the abundant labelled ImageNet images together with the unlabelled videos.
- We devise an effective video event recognition system based on the proposed semantic-visual knowledge, which act as intermediate-level semantic representation to handle the semantic gap problem.

The remainder of this paper is organized as follows. Section II discusses the related works and their limitations. Section III present the overview of our framework. Section IV constructs the semantic-visual knowledge base which contains event-centric concept relationship; while Section V proposes a robust transfer learning model which learns the noise-resistant concept classifiers. Section VI specifies the implementation details. Section VII lists the datasets we used and provides the results for all our experiments. Finally, we summarize our work in Section VIII.

II. RELATED WORK

In this section, we review the related works in three aspects. First, we review multimedia event recognition with respect to the semantic concept analysis. Second, in order to better understand the concept relationship, we probe into the semantic knowledge base and review how it is utilized in video content analysis. Finally, we briefly review recent works on concept learning technology in video domain.

A. Video Event Recognition

Low-level visual features extracted from images in the bag-of-word representation have been extensively used in video event detection. However, the low-level features are incapable of understanding the semantic structure presented in the long and complex video. Therefore, recent works explored the combination of low-level and high-level features. For example, Izadinia et al. [20] fused six different low-level features, such as SIFT, STIP, GIST, together with 62 activity concepts as
Fig. 2. Flowchart of the proposed video event recognition system. A list of events in the text descriptions and their video collections are given to the system. The text descriptions are used to produce a list of LUs from FrameNet, and then ImageNet is used to obtain relevant synsets as the event-related concepts by referring to the list of LUs and to obtain a set of training images for each obtained concept. The training images and the video collections are used to train the cross-media concept classifier. In the online process, for example, the aim is to recognize the event “Flash mob gathering.” This target event is augmented by the related frame “Aggregate.” The test video is fed into the concept classifier to obtain the high-level features and recognized by the event classifier.

high-level features. Ramanathan et al. [21] used SIFT, MFCC and other low-level features together with 13 roles and 46 actions. Sun et al. [22] fused the motion feature with 60 activity concepts. It seems that dense trajectory feature [23] is the single best feature, and other visual features complement each other.

Current methods for constructing the informative concept bank in favor of different events can be generalized as: (a) automatically select the distinguishing high-level features from video examples [7], [5], [13]; and (b) manually select the related high-level concepts [6], [11]. In the first method, the selections are based solely on the few positive video examples for a complex event, and thus may not conform to humans’ broad understanding of the event. On the other hand, the second method though more effective, is time consuming and subjective. As the concept corpus grows, users will need automated recommendation systems to assist in the selection of concepts.

As compared to these previous works, our aim is to design a robust and effective framework for automatically constructing the ideal composition of concepts for different events. The concept selection is based on several well-established knowledge bases.

B. Knowledge Base

Fortunately, many domain/commonsense knowledge bases have been encoded such as the WordNet [17] or Multimedia Ontologies [24]. Such knowledge bases provide a conceptual view of the domain at the schema level, which consist of concepts, concept properties, and relationships between concepts. Domain ontologies also provide a description of some specific application domain, for example, the ontology for soccer [25].

A recent work [26] designed a data-driven approach to build large-scale Visual Sentiment Ontology and establish semantic relations to keywords. It enables automatic sentiment analysis.

The external knowledge source has been used in text retrieval to enrich the query description. However, though it is easy to generate semantic relation between keywords, it remains challenging to incorporate the high-level semantic concepts into a visual retrieval task. The previous works of video annotation in sports and news domain use event-specific rules and templates to facilitate the recognition [27]. Some works use existing ontology such as WordNet to calculate the semantic similarity between concepts and construct the event representation [11], [6]. However, the concepts they used are text-based spoken words, rather than visual concepts. Compared to these methods, we not only explore the relation between keywords, but also visualize the event-related concepts.

C. Concept Learning in Video Domain

Due to the scarcity of the available labelled video data in video event recognition tasks, the concept classifiers are trained from auxiliary Web sources [8], [28], [13]. The conventional concept detection methods [10], [29], [30] built concept classifiers from labelled examples in one domain and assumed that the test data also comes from the same domain. However, in event recognition tasks [8], [10], [5], [4], the data distributions of the concept training domain and the video testing domain are normally different. The visualization of the feature differences in video and image domains can be found in Jeff’s work [31]. Recent works use domain adaption, or transfer learning, to propagate the knowledge learned from an auxiliary domain to a target domain. They can be roughly divided into two classes according to the auxiliary data: learning concepts from videos [28], [13], [32] and learning concepts from images [33], [34]. The tagged images are preferred in learning the concepts, as they contain more reliable tags and are of high quality. Duan [35] proposed a cross-domain kernel learning method minimizing the mismatch of data distribution between images and videos, but they still need the annotation information on the target video domain. Another machine learning framework called self-taught learning is proposed in [36] for using unlabeled data in supervised classification tasks. However, it does not consider the noise in labeled images. There are also concept learning methods [37], [38] which make use of the noise data. In this paper, we propose to learn the concept from labeled images and unlabeled videos considering the domain difference, and the method is noise-resistent.
III. FRAMEWORK OVERVIEW

The video events generalize what happens in the video content. It can be thought of a composition of various evidential description of scene, objects/people and activity. For example, the various evidences for the predefined event ‘working on a wood-working project’ are garage, saws and toolbox. An overview of our system is shown in Fig. 2. Given several predefined events and a set of training videos, we want to analyze both the event definition and the videos, so as to recognize the events depicted in unlabelled videos. We first analyze the events through a semantic-visual knowledge base construction, and then process the video by extracting the related concepts. Finally, we generate the video representation and do the event recognition task.

For the semantic-visual knowledge base construction, we make use of the knowledge base FrameNet, WordNet and ImageNet to generate the event related concept set \( \{C_1, \ldots C_n\} \), where all the concepts \( C_i \) are related to \( E_j \). The details are described in Section IV.

For the process of video, we segment a video into clips and key frames, and extract the low-level features. We also extract high-level features which are the scores of the presence of event related concepts. The concepts are described and augmented through semantic-visual knowledge base. As the general concept classifiers used in MED work [10], [4], [7] are not trained in video domain, we develop our cross-media concept learning method in Section V. However, the video samples alone are insufficient, so we transfer the knowledge from images to facilitate concept learning.

In order to generate the final video representation, we aggregate the key frames into video level, and employ the fusion strategy for high-level and low-level features. For specific, we evaluate three kinds of low-level features, including the visual and motion features. The event model is learnt through the simple and widely used linear SVM. For the test video, we extract the low-level features and apply the concept classifier on video frames to determine the high-level features. Finally, the fused high-level and low-level features are used to recognize events through the linear SVM.

IV. SEMANTIC-VISUAL KNOWLEDGE BASE

In order to recognize a predefined video event effectively, we need to identify concepts that are discriminative to that event. However, how to select such discriminative concepts is a non-trivial task. In most current approaches, the relationships between concepts, and especially the concept-event associations, are usually insufficiently formulated. It is often subjective and impossible to manually select the comprehensive event-centric concepts. On the other hand, though some semantic networks are available that capture a wide range of commonsense concepts and their relations, they are structured in a textual way, which lacks the visual knowledge. Thus, in this section, we propose an approach to find the most discriminative concepts for each event by using external knowledge sources and incorporate visual information to these concepts. We start from the event description, and introduce additional relevant concepts using the lexical knowledge base named FrameNet. In order to enrich the description in FrameNet to visual concepts, we make use of ImageNet. Finally, we generate a new knowledge base as shown in Fig. 3, which defines the event and concept relationship based on the semantic network together with the visual knowledge.

A. External Knowledge Sources

The FrameNet project was launched in UC Berkeley since 1997 [15]. FrameNet is constructed by computational linguists, who make great efforts in constructing the frame by finding words that fit the frames and annotating a large number of selected examples. For many years, researchers in NLP (Natural Language Processing) have been using FrameNet to support applications such as information extraction, machine translation, event recognition, and sentiment analysis. In FrameNet, the Frame is defined as a situation, scenario or event that involves an interaction (semantic relations) of its participants (semantic roles). FrameNet also defines the Lexical Units (LUs) as a set of words that evoke the event, where LUs resemble the event-centric concepts. The examples of some Frames and their participating Lexical Units are shown in Table I. A multimedia event can be mapped into the FrameNet framework by combining several related Frames together. For example, the event ‘wedding ceremony’ can be treated as the combination of the Frames Social_Event, Forming_Relationships and Personal_Relationship, because the word ‘wedding’ and ‘ceremony’ appear in these Frames. Thus all the possible concepts related to the multimedia event can be treated as the Lexical Units in these frames.

In addition, humans tend to use terms with varying generalizations to describe video events [4]. To be consistent with human’s understanding of the multimedia events, we use specific concepts as well as general concepts. The specific concepts
are more discriminative to one event, while the general concepts are frequent concepts that are shared by many events. Here we make use of WordNet, where terms are grouped into sets of cognitive synonyms (synsets), and are structured based on their hypernym/hyponym relations.

The above mentioned Lexical Units of FrameNet and synsets of WordNet express a distinct concept. However, they only provide text-based information. For a visual event recognition task, we need to associate concepts with images. In general, not all the concepts can be expressed or recognized visually. For example, words like ‘project’ and ‘purpose’ are hard to express in a visual way. Hence, we need to prune our semantic concepts to only visual concepts by using ImageNet, which is originated from WordNet and partially enriches WordNet with pictures.

B. Concept-Event Association

Based on the external knowledge sources, we generate our new Semantic-Visual Knowledge Base by using the procedure as outlined in Algorithm I. For each event, we generate its list of related concepts as follows. First, the words in each event description are used as the initial query terms to search for the event related Frames. We match the query word with the corresponding LUs in FrameNet. As we focus on the visual knowledge, we use only the noun word as query. For example, in the event ‘changing a vehicle tire’, we find that the noun query word ‘Vehicle’ is the LU of the Frame Vehicle and the noun word ‘Tire’ is the LU of the Frame Vehicle_Subpart. However, there are cases where one query word may be presented in multiple frames. The number of frames evoked by a query word is denoted as the spreading factor. If the spreading factor is larger than three, it means that the query word is ambiguous in identifying the events, and thus we ignore all those evoked frames. For example, the noun word ‘work’ is related to Frames such as Being_employed, Locale_by_use, Dimension, Work and Labor_product, thus it is hard to determine which frames are highly related. Second, we use the remaining evoked Frames and all their related LUs to search for more concepts. For example, in the Frame Vehicle, there are also LUs such as automobile, bikes, bus and car; while in the Frame Vehicle_Subpart, other LUs such as door, seat, tire, wheel and window can also be found.

All the LUs in these Frames then form the potential set of concepts. However, only visual concepts are useful for visual event detection task. Hence in the next step, we try to incorporate visual knowledge into these concepts. We refine the potential list of concepts by finding the corresponding noun synsets in WordNet, so as to generate the link to images in ImageNet. The simplest way to link is to compare the synset name with the LU words. For example, the synset ‘car tire’ contains the LU word ‘tire’. However, the synset name is not always a reliable indication for the same LU meaning, and hence we need to examine the context of a synset. The context of a target synset should either be consistent with the LU’s definition or has relation to the events. Synset gloss in WordNet contains a brief definition (‘gloss’) which illustrates the use of the synset members. Here we compare synset gloss and the LU’s definition to analyze the context. Moreover, to verify the relationship between synset and event, a possible concept pool (PCP) is constructed, which contains all the query words and all the related LUs. Note that we remove the stop words in the definition and gloss before the comparison is done. Next, we compare the WordNet synset’s gloss with all the possible concepts in the concept pool. If a match occurs in the comparison, the WordNet synset should be added into the concept set as the event-centric concept. Here is an example for ‘LU’s definition’ and ‘synset gloss’ to illustrate how the context of a synset is examined to select the concept set. The LU’s definition for ‘widow’ in FrameNet Personal_relationship is ‘a woman who has lost her husband by death and has not married again’. There is a synset ‘widow’ which simply contains this word. However, the synset gloss is ‘mostly black African weaverbird’. The gloss neither has any words contained in the LU’s definition nor in PCP, so LU’s definition and PCP are used to filter out unnecessary synsets. Another synset example ‘marriage bed’ is discovered by the LU ‘marriage’ but the gloss ‘the bed shared by a newly wedding couple’ does not contain the LU ‘marriage’, so there is a need to use other words to verify synset’s semantic consistency with the LU’s definition. Obliviously, the word ‘wedding’ is found in PCP, thus the synset is selected.

Due to the polysemy problem, we need to verify the reliability of concepts found in the concept set. The verification is based on word hierarchy. Specifically, in ImageNet, all the pictures are divided into nine classes of: ‘Plant’, ‘Geological formation’, ‘Natural object’, ‘Sport’, ‘Artifact’, ‘Fungus’, ‘Person’, ‘Animal’ and ‘Misc’. Here all the synsets can be traced back to one of these nine classes. Hence given a query word, we know which classes a related concept should not belong to, and we manually exclude certain classes for each query. Therefore, we eliminate some irrelevant synsets. For example, the LU ‘player’ which means ‘a machine that reads and decodes digital information’ is found in Frame ‘Gizmo’ by the query ‘appliance’. One of the related synset is ‘seeded player, seed’. The gloss definition ‘one of the outstanding players in a tournament’ contains the LU word ‘player’, so it is not filtered out through PCP. For this specific query, we manually select only the concepts related to ‘Artifact’ other than ‘Person’. Because the synset is the descendant of person, it is then removed.

**Algorithm I Event-centric concept set**

Input: query word set  
Output: concept set  
1: Find all queries related Frames  
2: Find all noun phrase LUs in related Frames  
3: Put LUs and queries in Possible Concept Pool (PCP)  
4: for LU in LUs do  
5: Find all WordNet synset (WS) names that contain LU  
6: for all WS do  
7: Compare WS gloss with LU’s definition  
8: Compare WS gloss with PCP  
9: if $WS \cap LU \neq \emptyset$ or $WS \cap PCP \neq \emptyset$ then  
10: Put WS in concept set  
11: end if  
12: end for  
13: end for  
14: Verify the concept set
Finally, we obtain the general concepts by adding the parent nodes of the above specific concepts. As our concepts come directly from the ImageNet, their parent nodes can be easily found by tracing back the ImageNet hierarchy. The description for EVVE dataset contains lots of places and names. We do not use these names as queries and only use the noun words such as concert, arrest, attack, so the generated concepts set is small compared to the MED11 dataset. The above procedure to obtain event-related concept set is shown in Fig. 3. Each concept can be related to one or many events. All the concepts are accompanied with pictures. An example of the four selected concepts ‘matchmaker’, ‘bridal gown’, ‘wedding’ and ‘domestic partner’ for event ‘wedding ceremony’ is shown in Fig. 4. We also note that, there still exist certain noises in the constructed concept set, for example, the synset ‘hope chest’ and ‘divorce lawyer’ are also selected by the query ‘wedding’.

V. CROSS-MEDIA CONCEPT LEARNING

Due to the scarcity of available positive examples of complex events and the absence of concept tags at key frame levels, many works [8], [28], [13] employ auxiliary image data to facilitate the training of concept classifier for video recognition. Furthermore, abundant Web images have merits as compared to Web videos for learning the concepts because they contain more reliable tags and the quality is relatively high. To specify, we use images from ImageNet. However, before images can be used to train concept classifiers for videos, we need to overcome the following two challenges. First, the concept detectors are partially built from image domain, while the test data is from video domain. The distributions of data from these two domains are different. Second, different levels of noise exist in the Web source. In the following Sub-Sections, we leverage a large number of labelled ImageNet images to facilitate the detection of concepts in video domain, and focus on dealing with the above challenges.

A. Minimizing the Domain Difference

The videos belonging to the concept related events are used as video domain data, because these videos have higher probabilities of relating to the training concept. Though the video frames are unlabeled data, they provide statistic information to bridge domain differences between the video and the image. We refer to the image domain as the auxiliary domain $D^A = \{ (x^A_i, y^A_i) \}_{i=1}^{N_A}$, and video domain as the target domain $D^T = \{ x^T_i \}_{i=1}^{N_T}$, where $x^A_i$ represents the visual feature vector of the $i$th image in $N_A$ images in training concept $C$, and $x^T_i$ represents the visual feature vector of the $i$th video key frame in $N_T$ video key frames, and $y^A_i \in \{ -1, 1 \}$ is the label of $x^A_i$ for a given concept. In cross-domain learning, the crucial step is to reduce the differences between auxiliary domain and target domain. Our strategy is to find an optimal higher dimensional space with minimum domain differences. Many works [35], [28] use the Maximum Mean Discrepancy to measure the difference between data distributions in the Reproducing Kernel Hilbert Space (RKHS)

$$\text{dist}(D^A, D^T) = \left\| \frac{1}{N_A} \sum_{i=1}^{N_A} \varphi(x^A_i) - \frac{1}{N_T} \sum_{i=1}^{N_T} \varphi(x^T_i) \right\|_H.$$  (1)

The samples are transferred into a higher dimensional space, through the non-linear feature mapping $\varphi(\cdot)$, where the dot product equals to a kernel function $\kappa(x, x') = \varphi(x)^T \varphi(x')$ (cf. Section VII-B for the details of kernel function). As showed in [35] and [28], the distance can be rewritten as

$$\text{dist}(D^A, D^T) = Tr(K_uS)$$

where $S = ss^T, Tr(\cdot)$ is the trace of the matrix.

$$s = \begin{bmatrix} 1 \quad \cdots \quad 1 \quad 1 \quad \cdots \quad 1 \end{bmatrix}^T$$  (2)

$$K_u = \begin{bmatrix} \kappa(x^A, x^A) & \kappa(x^A, x^T) \\ \kappa(x^T, x^A) & \kappa(x^T, x^T) \end{bmatrix}$$  (3)

$\kappa(x^A, x^T)$ refers to the kernel matrices over domain $A$ and domain $T$. In order to find the optimal kernel space, we optimize the linear combination weights of the $m$ base kernels $\kappa^{(1)}, \kappa^{(2)}, \ldots, \kappa^{(m)},$ defined on the samples from both the auxiliary and target domains, denoted as: $K_u = \sum_{k=1}^{m} u_k \kappa^{(k)}$. Then, by defining the combining weights $u = [u_1, \ldots, u_{mT}]$, where $u_k \geq 0, \sum u_k = 1$ and $p = [Tr(\kappa^{(1)}S), \ldots, Tr(\kappa^{(m)}S)]^T$, 1 can be rewritten as

$$\text{dist}(D^A, D^T) = u^T p.$$  (4)

B. Noise-Resistant Concept Classifier

We aim to learn a concept classifier by considering both the noise in the image labels as well as the above differences between the two domains. We define $x_{tr}^A \in \mathbb{R}^{N_A}$ as the training sample vector from $D^A$, and $v \in \mathbb{R}^{N_A}$ as the classification parameter. For a given sample $x_i$, the classifier is formulated as

$$f(x_i) = v^T K_u(x_{tr}^A, x_i)$$  (5)

$$K_u(x_{tr}^A, x_i) = \sum_{k=1}^{m} u_k \kappa^{(k)}(x_{tr}^A, x_i).$$  (6)
The original multiple kernel learning can be cast as the following problem:

$$\min_{\mathbf{u},\mathbf{v}} \lambda \| \mathbf{f} \|^2_{\mathbf{H}} + \frac{1}{N_A} \sum_{i=1}^{N_A} \ell(y_i^A, f(x_i^A)).$$

Here $\lambda$ is the regularization parameter. Together with the domain difference, the final optimization object function is

$$\min_{\mathbf{u},\mathbf{v}} \alpha (\mathbf{u}^T \mathbf{p}^T \mathbf{u}) + \lambda \| \mathbf{f} \|^2_{\mathbf{H}} + \frac{1}{N_A} \sum_{i=1}^{N_A} \ell(y_i^A, f(x_i^A)). \quad (7)$$

The parameter $\alpha$ controls the domain difference scale.

Now we consider the noise in the training set. To address the uncertainty arising from the noise in image labels, we use probabilistic methods proposed in [37] through several steps of relaxing the loss function $\ell(\cdot)$ in (7). We repeat their formulation below for ease of reading and making the paper self-contained. We denote $p_i \in [0,1]$ as the probability of each image to be rightly labelled. The parameter $\tau$ given the noise level $q \in [0, \frac{1}{2}]$ follows Hoeffding’s inequality that holds with a probability at least $1 - \epsilon$ given by

$$P = \{p_i \frac{1}{N_A} \sum_{i=1}^{N_A} p_i \leq 1 - q + \frac{\tau}{\sqrt{N_A}}\} \quad (8)$$

where $\tau = \sqrt{1/2 \ln(1/\epsilon)}$ according to the proposition in [37]. The final problem is

$$\min_{\mathbf{u},\mathbf{v}} \max_{p_i \in P} \alpha (\mathbf{u}^T \mathbf{p}^T \mathbf{u}) + \lambda \| \mathbf{f} \|^2_{\mathbf{H}} + \frac{1}{N_A} \sum_{i=1}^{N_A} (p_i + \tau) \ell(y_i^A, f(x_i^A)). \quad (9)$$

C. Optimization Solution

In this section we present our algorithm to minimize the problem in (9). As we incorporate the transfer learning part, the optimization process is different from the work in [37]. First, we rewrite the formulation by unfolding the hinge loss function with a parameter $\gamma_i \in [0,1]$. The loss function is

$$\max(0, 1 - y_i^A f(x_i^A)) = \max_{\gamma \in [0,1]} \gamma_i (1 - y_i^A f(x_i^A))$$

and we define $\{\beta_i = \gamma_i (p_i + \tau)\}_{i=1}^{N_A}$, so according to (8) the domain of $\beta_i$ is

$$B = \{\beta_i \in [0,1] + \tau \mid i \leq N_A (1 - q + \frac{\tau}{\sqrt{N_A}})\}.$$ 

Finally, the third term in (9) is rewritten as

$$\max_{p_i \in P} \frac{1}{N_A} \sum_{i=1}^{N_A} (p_i + \tau) \ell(y_i^A, f(x_i^A))$$

and defined $\{\beta_i = \gamma_i (p_i + \tau)\}_{i=1}^{N_A}$, so according to (8) the domain of $\beta_i$ is

$$B = \{\beta_i \in [0,1] + \tau \mid i \leq N_A (1 - q + \frac{\tau}{\sqrt{N_A}})\}.$$ 

Finally, the third term in (9) is rewritten as

$$\max_{\beta_i \in B} \frac{1}{N_A} \sum_{i=1}^{N_A} \beta_i (1 - y_i^A f(x_i^A)). \quad (10)$$

The problem is then formulated as

$$\min_{\mathbf{u},\mathbf{v}} \max_{\beta_i \in B} \alpha (\mathbf{u}^T \mathbf{p}^T \mathbf{u}) + \lambda \mathbf{v}^T \mathbf{K}_{\mathbf{u}} (\mathbf{x}_1^A, \mathbf{x}_2^A) \mathbf{v} + \frac{1}{N_A} \sum_{i=1}^{N_A} \beta_i (1 - y_i^A f(x_i^A)). \quad (11)$$

The above problem is solved in two steps. In the first step, we learn $\mathbf{v}$, and $\mathbf{u}$ is fixed. In the second step, we fix the learnt $\mathbf{v}$ to obtain $\mathbf{u}$. Referring to (6), with the fixed $\mathbf{u}$, we denote $L(\mathbf{v}, \beta)$ as the third term $\frac{1}{N_A} \sum_{i=1}^{N_A} \beta_i (1 - y_i^A f(x_i^A))$ in (11), and $R(\mathbf{v}) = \lambda \mathbf{v}^T \mathbf{K}_{\mathbf{u}} (\mathbf{x}_1^A, \mathbf{x}_2^A) \mathbf{v}$, Thus, we get the following problem:

$$\min_{\mathbf{u},\mathbf{v}} \max_{\beta_i \in B} \alpha (\mathbf{u}^T \mathbf{p}^T \mathbf{u}) + \lambda \mathbf{v}^T \mathbf{K}_{\mathbf{u}} (\mathbf{x}_1^A, \mathbf{x}_2^A) \mathbf{v} + \frac{1}{N_A} \sum_{i=1}^{N_A} \beta_i (1 - y_i^A f(x_i^A)). \quad (12)$$

Then we fix $\mathbf{v}$, and we denote $L(\mathbf{u}, \beta) = \frac{1}{N_A} \sum_{i=1}^{N_A} \beta_i (1 - y_i^A f(x_i^A))$ as the third term in (11), $R(\mathbf{u}) = \alpha (\mathbf{u}^T \mathbf{p}^T \mathbf{u}) + \lambda \mathbf{u}^T$. The problem is reduced to the following sub-problem:

$$\min_{\mathbf{u},\beta \in B} \max_{\mathbf{v}} R(\mathbf{u}) + L(\mathbf{u}, \beta). \quad (13)$$

We use the Primal Dual Prox method [37] to solve each of the above two optimization problems.

VI. FEATURES AND IMPLEMENTATION DETAILS

We extract key frames from videos at a uniform sampling rate (one frame in every five seconds). Three different kinds of low-level features are used in comparison. The first one is the combination of HOG and color moment, which has been found to be successful for object classification [39]. We utilize Fisher encoding to encode the visual features into a 73728D visual vector. For efficiency, we apply PCA to reduce the dimension to 2048D. The second is the SIFT feature. It has been widely used in multimedia event detection [2], [40]. The last one is a motion feature designed specifically for video event classification task [41]. We segment the video to 10 seconds short clips, with a step size of 5 seconds. Then, we extract dense trajectory features [23] on short clips, and apply Fisher Vector encoding under Gaussian Mixture Model. We denote this feature as DTFV. The original implementation is available online,1 and we use their codebook with the number of GMM clusters set to $K = 512$. The video-level representation is generated by average pooling through all the key frames and all clips.

We train the concept classifier using both the images from ImageNet and the video frames. To specify, we only use the videos in the event category to which the concepts are related. For example, we use the videos in ‘changing a vehicle tire’ and ‘Getting a vehicle unstuck’ to train the concept ‘self-propelled vehicle’. The concept classifier is trained from the low-level HOG features. Then, the concept classifier is applied to all the video frames to generate a high-level intermediate representation. The video-level concept scores are generated by average pooling through all scores on key frames.

---

As we extract both low-level features and high-level concept scores from videos, we need to further fuse them together to facilitate the video event recognition task. Therefore, we evaluate both early fusion and late fusion strategies. For early fusion, we directly concatenate all the concept scores with the low-level features. The concatenated features are fed to SVM for classification. For late fusion, we use the weighted average fusion method as proposed in [3]. In this method, high level concept scores and low level features are treated separately in two systems, and the threshold is automatically generated by minimizing the Normalized Detection Cost (minNDC) score. The confidence scores for the two system are denoted as $p_h$ and $p_l$. Given the threshold $T_h$, the weight for system $i$ is computed as

$$w_i = \begin{cases} \frac{T_h - p_i}{T_h - p_h}, & p_i < T_h \\ \frac{p_i - T_h}{p_l - T_h}, & p_i \geq T_h \end{cases}$$ (14)

The final score $P$ for a video is

$$P = \sum_i w_i p_i.$$ (15)

### VII. Experiment

#### A. Datasets

We evaluate our video event recognition method on two datasets. The first dataset is the TRECVID Multimedia Event Detection (MED11) [14] dataset. This dataset contains fifteen named event categories. As our proposed method depends on concepts relating to the events, we extensively use all the fifteen events. There are 2062 positive video samples in Event Kit. Following the protocol employed in works using the MED11 Event Kit [20], [22], [21], we use the same training (70%) and testing (30%) splits. The second dataset is a challenging Event Video (EVVE) dataset introduced in [42], which contains 166 hours of YouTube video data related to 13 events. Note that the EVVE dataset also includes distracter videos from similar but different events, which makes the recognition task much more challenging. The detailed statistics of the MED11 and EVVE dataset is shown in Table II and Table III.

2We got the splits by communicating with the authors.

### TABLE II

<table>
<thead>
<tr>
<th>Event</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Attempting boarding trick</td>
<td>113, 1.262</td>
<td>48, 639</td>
</tr>
<tr>
<td>2. Feeding animal</td>
<td>107, 1.268</td>
<td>55, 632</td>
</tr>
<tr>
<td>3. Landing fish</td>
<td>79, 1.296</td>
<td>43, 644</td>
</tr>
<tr>
<td>4. Wedding ceremony</td>
<td>76, 1.299</td>
<td>52, 635</td>
</tr>
<tr>
<td>5. Working wood working project</td>
<td>97, 1.278</td>
<td>46, 641</td>
</tr>
<tr>
<td>6. Birthday party</td>
<td>129, 1.246</td>
<td>44, 643</td>
</tr>
<tr>
<td>7. Changing a vehicle tire</td>
<td>74, 1.301</td>
<td>37, 650</td>
</tr>
<tr>
<td>8. Flash mob gathering</td>
<td>116, 1.259</td>
<td>57, 630</td>
</tr>
<tr>
<td>9. Getting a vehicle unstuck</td>
<td>87, 1.288</td>
<td>45, 642</td>
</tr>
<tr>
<td>10. Grooming an animal</td>
<td>84, 1.291</td>
<td>54, 633</td>
</tr>
<tr>
<td>11. Making a sandwich</td>
<td>80, 1.295</td>
<td>46, 641</td>
</tr>
<tr>
<td>12. Parade</td>
<td>95, 1.280</td>
<td>43, 644</td>
</tr>
<tr>
<td>13. Parkour</td>
<td>72, 1.303</td>
<td>40, 647</td>
</tr>
<tr>
<td>14. Repairing an appliance</td>
<td>82, 1.293</td>
<td>41, 646</td>
</tr>
<tr>
<td>15. Working on a sewing project</td>
<td>84, 1.291</td>
<td>36, 651</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>Event</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Austerity riots in Barcelona</td>
<td>15, 1.330</td>
<td>21, 1.325</td>
</tr>
<tr>
<td>2. Concert of Die toten Hosen, Rock am Ring,2012</td>
<td>33, 1.312</td>
<td>54, 1.312</td>
</tr>
<tr>
<td>3. Arrest of Dominique Strauss-Kahn</td>
<td>14, 1.331</td>
<td>12, 1.334</td>
</tr>
<tr>
<td>4. Egyptian revolution: Tahrir Square demonstrations</td>
<td>57, 1.288</td>
<td>49, 1.297</td>
</tr>
<tr>
<td>5. Concert of Johnny Hallyday stade de France,2012</td>
<td>115, 1.230</td>
<td>124, 1.222</td>
</tr>
<tr>
<td>6. Kate William wedding</td>
<td>57, 1.258</td>
<td>58, 1.288</td>
</tr>
<tr>
<td>7. Bomb attack in the main square of Marrakech</td>
<td>6, 1.339</td>
<td>6, 1.340</td>
</tr>
<tr>
<td>8. Concert of Madonna in Rome,2012</td>
<td>71, 1.274</td>
<td>65, 1.281</td>
</tr>
<tr>
<td>9. Presidential victory speech of Barack Obama,2008</td>
<td>18, 1.327</td>
<td>17, 1.329</td>
</tr>
<tr>
<td>10. Concert of Shakira in kiev,2011</td>
<td>21, 1.324</td>
<td>20, 1.316</td>
</tr>
<tr>
<td>11. Eruption of Stromboli geyser in Iceland</td>
<td>314, 1.031</td>
<td>301, 1.045</td>
</tr>
<tr>
<td>12. Major autumn flood in Thailand</td>
<td>88, 1.257</td>
<td>87, 1.259</td>
</tr>
<tr>
<td>13. Jurassic park ride in universal studios theme park</td>
<td>29, 1.316</td>
<td>33, 1.313</td>
</tr>
</tbody>
</table>

Two evaluation metrics are used. The first is Minimum NDC (MinNDC), which is the official metric defined by NIST [14]. For each event E, we have

$$NDC(E) = \frac{0.999 \times P_F(A_E) + 0.08 \times P_M(D_E)}{0.08}$$

where $P_M$ is the missed detection probability and $P_F$ is the false alarm probability. The MinNDC is introduced to ignore the decision threshold. Lower MinNDC indicates better performance. The second metric is Average Precision (AP). The videos are ranked in order of the score $P$. Higher AP indicates better performance.

#### B. Concept Classifier Performance

In this subsection, we evaluate the performance of the proposed cross-media concept classifier. The evaluation is done on the target video domain. We manually labelled 4,008 video key frames in the event of ‘changing a vehicle tire’. For these 4,008 frames, each frame has ten labels (0 or 1 for each label), corresponding to the ten concepts listed in Table IV. We do not include general concepts such as ‘self-propelled vehicle’, ‘wheeled vehicle’ in this concept evaluation set. The detailed statistics of this dataset are shown in Table IV. The positive images are from the exact category of the ImageNet, while the negative images are selected from the other nine categories. Note that the number of the negative instances is double the size of the positive instances.
As mentioned in Section V-B, the noise part derives from the technique described in [37], and we inherit their settings; For example, we use the 10-Gaussian-kernels as the base kernels. As we use the MKL framework, the proposed method computes a set of kernel matrices for each sample, and learns a linear combination of these base kernels. The major computational cost is on the expensive computation of kernel matrices and the optimization on linear combination. In optimization, $\tau$ and $\eta$ are absorbed in the boundary $B$ of $\beta$. We use $N_\alpha$ as the boundary for $\beta$. The parameter $\alpha$ controls the domain difference scale. We fix $\lambda = 1$ and tune $\alpha$, and test the concept classification result (mean AP) on the target domain to determine $\alpha$.

The auxiliary images are collected from ImageNet. We evaluate different $\alpha$ within $\{0.01, 0.1, 1, 10, 100, 1000\}$. The performances on video key frame set are shown in Fig. 5. As $\alpha$ increases, the performance is improved, but it will decrease when alpha exceeds 100. As shown in (9), $\alpha$ balances the weight between the domain difference and the original classification loss. Larger $\alpha$ emphasizes more on minimizing the domain difference, however, the penalty on the classification loss will relatively reduce. In summary, the results are improved by increasing $\alpha$ within a range.

Fig. 6 visualizes the low-level HOG features for the ten concepts in both the video and image domains, which suggests the domain bias. We use the first two components after PCA to find the 2-dimensional embedding of the high-dimensional feature space. Note that we use all the training samples listed in Table IV to plot this figure. Each sample is plotted as a point colored depending on their concept category. The ten colors represent the ten concepts as listed in Table IV. The left one in Fig. 6 shows the bivariate distribution of HOG features in image domain and the right one shows the distribution in video domain. As shown in Fig. 6, the average Y value of the blue points in video domain is smaller than that in image domain. The shapes of the feature distributions for the same concept are different between two domains. Moreover, in the video domain, the features of different concepts tend to overlap, as opposed to the image domain, thus showing the need for transfer learning methods.

Fig. 7 evaluates the contribution of different kinds of training data and compares our concept learning method to the transfer learning and self-taught learning methods. As shown in Fig. 7, the seven methods use training data from different domains and employ different learning methods, but test on the same video key frame set. We use the same low-level HOG feature for all these methods. For the first method IMAGE, we only use the labeled images to train the model. For the second method, we only use the labeled video frames in training without any image, and it is denoted as VIDEO. For the third method, we train the model with labeled images together with labeled video frames, and it is denoted as IMAGE+VIDEO. Note that in all the above methods, we use SVM to learn the concept without considering the domain difference between the video and image. The fourth and fifth methods A-MKL [28] and DTSVM [35] are cross-domain kernel learning methods which minimize the domain difference in optimization. The implementations of these two methods are available online. The sixth method Self-taught [36] is based on sparse coding and learns from both labeled and unlabeled domain. The seventh method is our noise-resistant learning method DTnoise. We categorize the above seven methods to three groups according to the usage of different training data:

- single domain: IMAGE, VIDEO;
- labeled image + labeled video frame: IMAGE+VIDEO, A-MKL, DTSVM;
- labeled image + unlabeled video frame: Self-taught, DTnoise.


In comparison of methods IMAGE and VIDEO, the performance of using video frames is better than using images for learning concepts such as 'wagon' and 'jeep', though the positive instances appeared in the video frames are less than that in images. This result suggests the model trained from image domain is not good enough to predict concepts in video domain. It also demonstrates the difference between image domain and video domain. When we use both the labeled images and video frames in training (IMAGE+VIDEO), the overall performance is promoted. The training data are from diverse domains, and both have annotations, so the classification performance is improved. The mean AP (mAP) over all the evaluated concepts is 0.430 for IMAGE+VIDEO method, compared to 0.387 for IMAGE method and 0.395 for VIDEO method. Furthermore, the results of A-MKL and DTSVM indicate that considering the difference between image and video domain further boosts the concept learning performance. We come to a conclusion that exploring images and video frames together will generate a better result. However, the annotation information on video frames is limited, so we cannot rely on frames to learn the concept model. The Self-taught method and our method DTnoise are capable of incorporating the unlabeled video frames in learning the concepts. The Self-taught uses image data to guess the video’s label, while our method only minimizes the domain difference without the guess of concept in the training videos. The mAP over all the evaluated concepts is 0.401 for our method and 0.396 for the Self-taught method. These two methods outperform that of the IMAGE method by around 3%. It suggests that the incorporation of unlabeled video frames is useful. Note that for the concept 'race car', the Self-taught method performs much better. The possible reason is the image data seems providing more discriminative information than the video data for this concept, considering the fact that both the VIDEO and IMAGE+VIDEO methods are not good. Therefore, for this concept, the methods that rely more on the image data outperform our method. It is worth noting that 'car tire' and 'radial tire' are similar concepts, but the images used in the training are from different ImageNet synsets. Most of the images from the synset 'radial tire' capture the close-up tires in a constrained environment with clear background; hence, the appearances of these 'radial tire' images are much more different from the tires in the videos. In our method DTnoise, the AP for 'radial tire' is 0.347, while that for the non-transfer learning method (IMAGE) is 0.335. The comparison suggests that the improvement of our transfer learning method is more apparent when there is a larger domain difference. Moreover, we notice that in ImageNet the quality of synsets is different. For example, the synset 'car door' in ImageNet contains lots of hatchback of a car instead of the real car door. Compared to the Self-taught method, our noise-resistant method achieves good result on such concepts, which again suggests its reliability and robustness.

C. Effects of Event-Centric Concept Set

Finally, we select 187 event related concepts for MED11 dataset [14] and 101 concepts for EVVE dataset [42]. We conduct four experiments to demonstrate the effectiveness of the event-centric concept set in recognizing video events. The first experiment shows the performance gain by using our high-level concepts. The second experiment evaluates results of using only the high-level concepts. The third experiment validates our selected concepts as compared to other concepts set. The fourth experiment gives examples to show the contribution of the groups of concept in recognizing the events.

In the first experiment, the results of using only the high-level features is evaluated. We analyze the contribution of general and specific concepts. The results on the MED11 dataset are illustrated in Fig. 8(b). Note that only the high-level concept scores are used without fusing with the low-level features. For different events, the specific and general concepts make different contributions. For example, the event ‘Landing fish’ contains different kinds of animals, thus the general concept like ‘animal’ will generate a concise summarization of the event; nonetheless, for the event ‘wedding ceremony’, the general concept like ‘person’ is less discriminative, hence for this event, specific concepts are more useful. In general, for most events, the combination of specific and general concepts always performs the best.

The second experiment evaluates the contribution of high-level (HL) concepts as compared to three kinds of low-level features: HOG, SIFT and DTFV. The results are shown in Table V. To specify, DTFV+HL refers to the method using labeled image to train the concept classifiers rather than our proposed concept learning methods. In all the three comparisons the performance is improved when the low-level features are fused with the event-specific high-level features. Note that the high-level concepts include both specific and general concepts, and early fusion is adopted. Results in column six show that the motion based features (DTFV) obtain better

![Fig. 8. Effects of event-centric concepts on MED11: (a) comparisons of early and late fusions [HOG features, HOG features + high-level features (all concepts)], MED11, and (b) comparisons of specific, general, and both concepts (only high-level features), MED11.](Image 303x395 to 553x508)

![Fig. 8. Effects of event-centric concepts on MED11: (a) comparisons of early and late fusions [HOG features, HOG features + high-level features (all concepts)], MED11, and (b) comparisons of specific, general, and both concepts (only high-level features), MED11.](Image 305x620 to 550x725)
results than other low-level features, this is consistent with the results in works [2], [3] in evaluating the low-level features. When combined with high-level concepts, the mean average precision can be further improved by 2.7%, which indicates that the event-centric concepts have complementary performance. Meanwhile, the comparison between DTFV+HL’ and DTFV+HL shows that the cross-media model exerts positive effects on the concept learning, which also helps to guarantee the superior performance. Though the concept detector is also trained from the low-level HOG feature, the overall mean average precision of our approach exceeds that of HOG feature by around 9% as shown in the first two columns. It indicates that the high-level concept scores can be seen as nonlinear projections of original feature space, which provide discriminative information for recognizing the video event. For the event with large intra-class variations on low-level features such as the ‘birthday party’, the improvement is more significant. This shows that the event-specific concepts are able to complement the low-level features, and in some degree, they help narrow down the semantic gap between the high-level complex events and low-level visual features. The results in Fig. 8(a) show that early fusion and late fusion methods generate comparable results. In the fusion period we use both the specific and general concepts. In the following experiments, we will use only the early fusion method.

In the third experiment, we compare our event-centric concept set to one of the state-of-the-art concept set, namely Object Bank [9], [8], which consists of 177 most frequent objects (OB concepts). For our approach, we generate 187 event-centric concepts from the knowledge base for MED11 and 101 concepts for EVVE. 5 For fair comparison, we use the same features and the same concept learning methods. As the 177 concepts in Object Bank are from ImageNet as well, the training samples are from the same source. Therefore, the only difference is the concept set. To specify, we consistently use our cross-media concept learning model to train all the concept classifiers (177 for OB concepts, 187 for MED11 and 101 for EVVE), and we adopt HOG features as the low-level features. Fig. 9(a) and Fig. 9(b) report the performance on MED11 and EVVE, respectively. As we can see, the mAP of our event-centric concepts (i.e., 0.616 for MED11 and 0.784 for EVVE) outperforms that of the OB concepts (i.e., 0.525 for MED11 and 0.724 for EVVE) by around 12%. Besides, in most cases, our approach can achieve better AP on each individual event than the use of OB objects. This is because as compared to the much more general semantics in Object Bank, our proposed event-centric concept set is augmented with more comprehensive and specific semantics from the semantic-visual knowledge base, thereby able to induce more discriminative power for facilitating the video event recognition task.

Further, we give some examples to analyze the groups of concepts for some events. We adopt a simple way to calculate the contribution of some discovered concepts in recognition of the events. In Fig. 10, there are two test videos from event ‘wedding ceremony’ and ‘birthday party’. When using all the 187 high-level concepts to recognize these two videos, they get positive scores for their corresponding event (1.41 and 1.91). Then, we exclude one related concept from the concept classifier and use the other 186 concepts to evaluate how the video score changes. For example, we exclude the concepts ‘bride’, ‘celebrant, celebrator’, ‘groom’, ‘matchmaker, marriage broker’ and ‘cake_mix’ one by one. The increase or decrease of the video score indicates the contribution of each concept. As shown in Fig. 10, without the concepts ‘bride’ and ‘groom’, the score of the first video is decreased, so these two concepts have more contribution to the event ‘wedding ceremony’. As for the video belonging to ‘birthday party’, the concept ‘cake_mix’ contributes the most. The concept ‘celebrant, celebrator’ is related to the event ‘wedding ceremony’, and it also benefits the recognition of birthday event. As the high-level features are imperfect, some interpretation may disagree with human’s prior knowledge. However, the experiment demonstrates that our semantic framework is capable of interpreting the final result.

D. Performance on Video Event Recognition

In this subsection, we compare our approach to state-of-the-art techniques in video event recognition. Izadinia et al. [20] proposed to use latent graphical model to capture the co-occurrence between concepts. Sun et al. [22] focused on exploiting the underlying temporal structure shared by concepts. Ramanathan et al. [21] jointly modeled the image and text data. Following the same protocol as suggested in [20]–[22], we
TABLE VI
COMPARISONS WITH METHODS USING OTHER CONCEPT SETS [DTFV + HIGH-LEVEL FEATURES (EARLY FUSION, ALL CONCEPTS)] AND LOW-LEVEL FEATURES (SIFT/MFCC/DTFV/STIP) + HIGH-LEVEL FEATURES (ROLES/ACTIONS), MED11

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.757</td>
<td>0.882</td>
<td>0.840</td>
<td>0.924</td>
</tr>
<tr>
<td>2</td>
<td>0.565</td>
<td>0.461</td>
<td>0.460</td>
<td>0.665</td>
</tr>
<tr>
<td>3</td>
<td>0.722</td>
<td>0.789</td>
<td>0.659</td>
<td>0.803</td>
</tr>
<tr>
<td>4</td>
<td>0.675</td>
<td>0.811</td>
<td>0.787</td>
<td>0.821</td>
</tr>
<tr>
<td>5</td>
<td>0.653</td>
<td>0.623</td>
<td>0.357</td>
<td>0.728</td>
</tr>
<tr>
<td>6</td>
<td>0.782</td>
<td>0.814</td>
<td>0.901</td>
<td>0.723</td>
</tr>
<tr>
<td>7</td>
<td>0.477</td>
<td><strong>0.518</strong></td>
<td>0.501</td>
<td>0.509</td>
</tr>
<tr>
<td>8</td>
<td>0.919</td>
<td>0.877</td>
<td>0.924</td>
<td>0.962</td>
</tr>
<tr>
<td>9</td>
<td>0.691</td>
<td>0.772</td>
<td>0.617</td>
<td>0.797</td>
</tr>
<tr>
<td>10</td>
<td>0.51</td>
<td><strong>0.634</strong></td>
<td>0.542</td>
<td>0.557</td>
</tr>
<tr>
<td>11</td>
<td>0.419</td>
<td>0.524</td>
<td>0.570</td>
<td>0.567</td>
</tr>
<tr>
<td>12</td>
<td>0.724</td>
<td>0.77</td>
<td>0.734</td>
<td>0.776</td>
</tr>
<tr>
<td>13</td>
<td>0.664</td>
<td>0.89</td>
<td>0.614</td>
<td>0.898</td>
</tr>
<tr>
<td>14</td>
<td>0.782</td>
<td>0.634</td>
<td><strong>0.784</strong></td>
<td>0.762</td>
</tr>
<tr>
<td>15</td>
<td>0.575</td>
<td>0.621</td>
<td><strong>0.669</strong></td>
<td>0.509</td>
</tr>
<tr>
<td>mAP</td>
<td>0.661</td>
<td>0.708</td>
<td>0.664</td>
<td>0.733</td>
</tr>
</tbody>
</table>

report our performance on fifteen events shown in Table VI. Columns 1,2 and 3 are quotes from the best results published in [20]–[22], respectively. All these methods make use of the activity concepts and model the correlations. Their concept set is either from human annotation on EventKit or standard human action dataset. As low-level motion features are used in all these three methods, we fuse our high-level features with the Dense Trajectory feature (DTFV+HL) to make fair comparison. We can see that our results achieve the best mAP, with between 3.5% to 10.9% higher performance over the others. According to the fourth columns of Table VI, our method achieves the best performance in 9 out of 15 events.

Further, we compare our approach to methods employing the static object concepts, namely Object Bank (OB) [9] and Concept Vocabulary [4] on both MED11 and EVVE. As aforementioned, our method uses early fusion strategy to fuse the 2,048D low-level HOG feature and the 187D high-level concepts as representation. For Object Bank, following the work in [8], we utilize the 44,604D feature generated by 177 object filters with max pooling. Therefore, it is different from the experiment in Section VII-C. Fig. 11 and Fig. 12 show the performance comparison based on the two evaluation metrics: MinNDC and AP. In most cases, our approach is able to outperform OB on the detailed AP on each event; with the overall mAP of our approach exceeds that of OB by around 16% on MED11 and 32% on EVVE. This phenomenon clearly indicates that the automatically constructed semantic-visual knowledge base indeed provides more useful semantics and visual resources for enhancing video event recognition. Compared to MED11, the EVVE dataset contains few positive instances; therefore, without informative prior knowledge into video event, the performance of OB method on EVVE dataset is much degraded due to the limited discriminative power of the general OB concepts.

Finally, we compare our method to that based on the 1,346 concept vocabulary as suggested in [4]. These 1,346 concepts are derived from the Semantic Indexing task and ImageNet Large Scale Visual Recognition Challenge, and are widely used in video event detection [7], [13], [5]. We exploit their provided training and testing data (MED11*), and use only the MED11 data for the evaluation. As can be observed in Fig. 13, in most cases, our method is able to achieve superior or comparable performance as compared to the use of 1,346 concept vocabulary. Though the 1,346 concept vocabulary contains more concepts, the presence of those concepts seldom changes the relevance of the videos and specific events, and thus the performance degrades especially for events with large intra-class variation such as ‘wedding ceremony’ and ‘parade’. Particularly, as the number of positive instances is limited, our event-centric concepts are better in capturing the specific pattern of the event. For the event ‘getting a vehicle unstuck’, our method generates worse AP but better MinNDC. This means, though the classification performances are evaluated to be worse than the concept vocabulary, our method has less errors when the miss-to-false-alarm-probability ratio is 12:5:1. For events such as the ‘making a sandwich’ and ‘working on
a sewing project', our performance is relatively unsatisfactory due to the lack of action modelling.

VIII. CONCLUSION

In this paper, we designed an effective system for video event recognition, based on the automatically constructed semantic-visual knowledge base. First, we encoded rich event-centric concepts and their relationships from the well-established lexical database FrameNet and WordNet. Second, we proposed a robust transfer learning model to learn the noise-resistant concept classifiers from ImageNet. Finally, we utilized the event-centric semantic concepts encoded in the knowledge base as the intermediate-level event representation. Extensive experiments on various real-world video datasets demonstrated that our method is able to generate discriminative concept set for event recognition as compared to the state-of-the-art approaches. In the future, we will incorporate action concepts based on the verbs in FrameNet and make use of Web images from various sources to facilitate the recognition.

REFERENCES

Xishan Zhang received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 2011, and is currently working toward the Ph.D. degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, Beijing, China.

Her current research interests include multimedia analysis and logo detection.

Jintao Li received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1989.

He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia technology, virtual reality technology, and pervasive computing.

Yang Yang received the B.S. degree from Jilin University, Changchun, China, in 2006, the M.E. degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from the University of Queensland, Brisbane, Australia, in 2013.

He was a Research Fellow with the School of Computing, National University of Singapore, Singapore. He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia information retrieval, social media analysis, and machine learning.

Hanwang Zhang received the B.Eng (Hons.) degree in computer science from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2014.

His research interests include multimedia and computer vision, focusing on developing techniques for efficient search and recognition in image contents.

Yongdong Zhang (AM’08–SM’13) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002.

He is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored over 100 refereed journal and conference papers. His current research interests include multimedia content analysis and understanding, multimedia content security, video encoding, and streaming media technology.

Dr. Zhang serves as an Editorial Board Member of the Multimedia Systems Journal and Neurocomputing. He was a recipient of the Best Paper Award in PCM 2013, ICIMCS 2013, and ICME 2010, and the Best Paper Candidate in ICME 2011.

Tat-Seng Chua joined the National University of Singapore, Singapore, in 1983, and spent three years as a Research Staff Member with the Institute of Systems Science, National University of Singapore. He was the Acting and Founding Dean of the School of Computing, National University of Singapore, from 1998 to 2000. He is currently the KIT/HTCT Chair Professor with the School of Computing, National University of Singapore. His research interests include multimedia information retrieval, multimedia question answering, and the analysis and structuring of user-generated contents.

Dr. Chua has organized and served as a Program Committee Member of numerous international conferences in the areas of computer graphics, multimedia, and text processing. He was the Conference Co-Chair of ACM Multimedia in 2005, the Conference on Image and Video Retrieval in 2005 and ACM SIGIR in 2008, and the Technical PC Co-Chair of SIGIR in 2010. He serves on the editorial boards of the ACM Transactions of Information Systems, Foundation and Trends in Information Retrieval, The Visual Computer, and Multimedia Tools and Applications. He is on the Steering Committees of the International Conference on Multimedia Retrieval, Computer Graphics International, and Multimedia Modeling Conference Series. He serves as a member of international review panels of two large-scale research projects in Europe. He is the Independent Director of two listed companies in Singapore.