Machine Vision and Applications (2014) 25:49–69
DOI 10.1007/s00138-013-0525-x

SPECIAL ISSUE PAPER

Multimedia event detection with multimodal feature fusion and temporal concept localization

Sangmin Oh · Scott McCloskey · Ilseo Kim · Arash Vahdat · Kevin J. Cannons · Hossein Hajimirsadeghi · Greg Mori · A. G. Amitha Perera · Megha Pandey · Jason J. Corso

Received: 11 January 2013 / Revised: 23 May 2013 / Accepted: 30 May 2013 / Published online: 16 July 2013

Abstract We present a system for multimedia event detection. The developed system characterizes complex multimedia events based on a large array of multimodal features, and classifies unseen videos by effectively fusing diverse responses. We present three major technical innovations. First, we explore novel visual and audio features across multiple semantic granularities, including building, often in an unsupervised manner, mid-level and high-level features upon low-level features to enable semantic understanding. Second, we show a novel Latent SVM model which learns and localizes discriminative high-level concepts in cluttered video sequences. In addition to improving detection accuracy beyond existing approaches, it enables a unique summary for every retrieval by its use of high-level concepts and temporal evidence localization. The resulting summary provides some transparency into *why* the system classified the video as it did. Finally, we present novel fusion learning algorithms and our methodology to improve fusion learning under limited training data condition. Thorough evaluation on a large TRECVID MED 2011 dataset showcases the benefits of the presented system.

S. Oh (✉) · I. Kim · A. G. A. Perera · M. Pandey
Kitware Inc., Clifton Park, New York, USA
e-mail: sangmin.oh@kitware.com

S. McCloskey
Honeywell Labs, Minneapolis, USA

A. Vahdat · K. J. Cannons · H. Hajimirsadeghi · G. Mori
School of Computing Science,
Simon Fraser University, Burnaby, Canada

J. J. Corso
Department of Computer Science and Engineering,
SUNY at Buffalo, Buffalo, USA

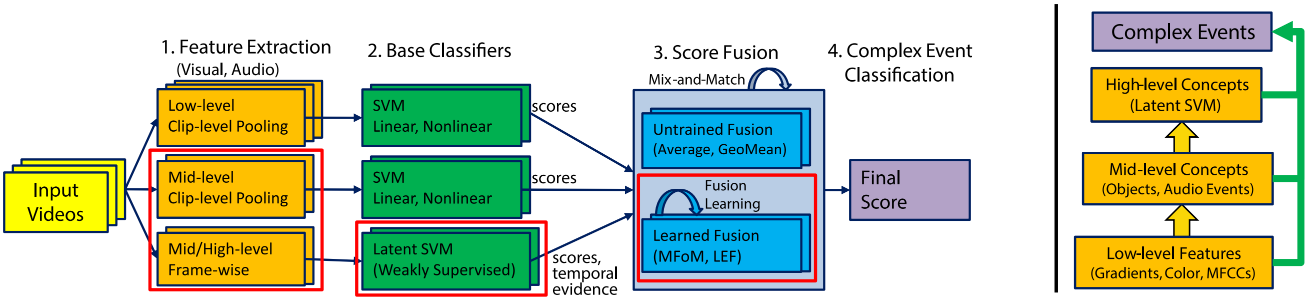
Keywords Multimedia · Classification · Machine learning · Fusion

1 Introduction

Multimedia content is being produced and shared through the Internet (e.g., YouTube) at an unprecedented pace in recent years. For example, just on YouTube, video data is currently being uploaded at the rate of approximately 30 million hours a year. Accordingly, the need for automatic tools that organize and retrieve videos has become crucial more than ever.

In this paper, we examine the task of multimedia event detection (MED), where the goal is detecting or classifying video clips by the main event occurring in the clip. In particular, we focus on high-level events such as 'making a sandwich', 'parkour', and 'grooming an animal', where the event is defined by complex collection of elements including people, objects, actions, sounds, scenes, and their spatial/temporal relationships.

Automatic understanding of visual content in unconstrained Internet videos is a very challenging task, particularly because they contain very diverse content. Even videos from the same event class (e.g., 'making a sandwich') exhibit significant intra-class variation: they are captured under a variety of camera conditions (e.g., pixel quality and motion); they are of highly variable duration and often heavily edited (e.g., shot stitching and added captions); they are typically not choreographed and do not follow a particular structure; the evidence indicating the presence of a particular high-level event typically occurs during specific segments of a video and is cluttered amid extraneous content introduced by diverse editing.



(a) System Architecture. **(b)** Concept Hierarchy

Fig. 1 **a** Overview of system architecture. Modules corresponding to major technical innovations are marked (*red boxes*). Refer to Sect. 3 for details. **b** Hierarchy of features across multiple granularities. Higher-level features are built upon the lower-level ones, and jointly characterize complex events. High-level concepts are utilized by Latent SVM model (see Sect. 5.2) (color figure online)

Due to the importance of the MED task, many state-of-the-art systems have been developed and reported by world-class researchers recently [18,21,23,30,39,48–50,59]. The main observations drawn from the success of these systems are consistent. Most systems extract a large number of multimodal features from both visual and audio signals. The features include low-level features such as SIFT, HoG, and MFCCs which provide significant portion of performance for MED. In addition, the use of pre-defined concept detectors for objects (e.g., person and cars) and actions (e.g., jump and punch) are also frequently incorporated. Then, the videos are classified by fusing features where score fusion is often used to combine scores independently computed from different subsets of features.

In this paper, we present a video event classification system which addresses various core challenges for MED task. Overall, our system is designed to incorporate many of the above-mentioned success principles in its architecture. Figure 1a shows our system architecture: it extracts large array of multimodal features at multiple granularities; multiple base classifiers are built from different subsets of features; finally, score fusion combines multiple scores and improves the MED performance. In particular, our work incorporates novel developments into the system, which can be summarized into three major contributions (highlighted in red boxes in Fig. 1a).

First, our work developed and incorporated novel features at diverse granularities, aiming to provide more semantic understanding capability into our system. The hierarchy of granularities and their relationships towards the modeling of complex events is illustrated in Fig. 1b.[1] For example, our work explores the use of *mid-level* concept features, which are detected based on low-level features. The mid-level features include (1) visual Object Bank (OB) [34] which provides 177 object detectors for pre-defined classes, and (2) a large array of audio event detectors [6] based on low-level MFCC features. Furthermore, this work developed novel *high-level* visual scene concepts which are discovered from the OB response distributions in an *unsupervised* manner, which are effective in discovering semantically consistent visual scenes. Note that this is distinct from more conventional approaches based on using pre-defined concept classes. Finally, all these features are *jointly* used to model complex events, where the strengths and limitations of each are exploited.

Second, we present a novel approach for weakly supervised semantic concept selection based on a Latent SVM (LSVM) formulation [13,56], which not only improves classification but also provides a novel scheme for *temporal evidence localization*. This approach addresses the temporal clutter issue mentioned earlier, where salient audio-visual evidence is buried amid less relevant content. In particular, this approach effectively utilizes "high-level" concepts mentioned earlier. For example, mid-level concepts correspond to general objects (e.g., tree, hat, computer, people) that may appear across a wide range of events. In contrast, "high-level" concepts capture a higher level of semantic information that is more specific to the event occurring in a particular video (e.g., skateboarding in garage, surfing on water). In detail, under the proposed approach, high-level concept classifiers are trained in an unsupervised fashion, without requiring expensive manual annotations. Then, our approach automatically identifies the classes and temporal locations of discriminative high-level concepts for a given target event class, which are both treated as hidden variables, effectively solving a weakly-supervised learning problem. Specifically, we show how the model parameters can be learned using Latent SVMs [13,56], a max-margin discriminative method that provides the ability to model unobserved variables during training. Through evaluation, the proposed method shows substantially superior performance beyond conventional approaches. Furthermore, we show that the capability to localize salient video segments provide a novel and effective summary of retrieval results.

[1] Note that the use of the terms, "mid-level" and "high-level" may be different from other work.

Third, we present two novel approaches [26,35] to *learn* score fusion functions which combine scores from different base classifiers. These fusion learning algorithms provide strengths under different circumstances and potentially boost performance beyond existing approaches. In this work, they are thoroughly evaluated to combine scores from a large array of 21 base classifiers, and their performance compared with existing approaches. In addition, we present our methodology to generate cross-validation splits to learn fusion classifiers, which is designed to maximize the utility of limited training data during fusion learning. Finally, we present the rationale for the *mix-and-match* fusion methodology incorporated in our system. For example, our evaluations indicate that there is no clear winner among the compared fusion methods and performance gaps between different methods can be more than trivial. Accordingly, multiple fusion approaches are tried per target event class during training, and our system selects the best approach case by case.

The proposed event classification system has been extensively applied to the challenging TRECVID MED 2011 dataset [42], which consists of training exemplars for 10 high-level event classes (listed in Table 2) and provides a large test data archive of about 32 K video clips. The evaluation results of our full system as well as the results from modular evaluation of key algorithmic components highlight the benefits of the proposed system.

The rest of the paper is organized as follows: Sect. 2 reviews the related work in multimedia event detection. Section 3 describes the overall system architecture. In Sect. 4, a set of features incorporated in our system are introduced. In Sect. 5, diverse approaches to build base classifiers are described, where Sect. 5.2 particularly focuses on the newly developed Latent SVM formulation. Sect. 6 presents our framework for fusion learning, along with two new score fusion learning algorithms. Finally, Sect. 7 presents the evaluation results along with various discussions on various interesting observations.

2 Related work

2.1 Content-based video retrieval

Previous work in the area of video detection and retrieval primarily considers the task of query-based video retrieval using concept detectors. The problem of video classification can be viewed as a subproblem of query-based video retrieval, where the query is pre-defined. A recent and thorough survey of content-based video retrieval techniques can be found in [19]. We discuss some of the closely related video retrieval/classification methods here.

In the past, research on video retrieval has largely focused on the use of pre-specified lexicons of concepts such as LSCOM [18] and MediaMill [49]. These lexicons typically cover a wide range of concepts including scenes, people, objects, activities, and events at different levels of abstraction. These lexicons, however, do not account for the fact that the appearance of a concept can vary from one event class to another. For example, people may dress differently depending on the environment they are in (e.g., snow vs. beach). These differences become more noticeable when using concepts at higher levels of abstraction because of an increased semantic gap [14]. This implies that concepts which represent a higher-level of abstraction depend more on the event category and consequently have limited reusability among different events.

A second challenge when using human specified lexicons is that the concepts may not reflect the true corpus found in the data. To address this challenge, Bao et al. [4] proposed the use of a bipartite graph propagation model to incorporate both human specified concepts as well as implicit concepts extracted by Latent Dirichlet Allocation for video retrieval. Another approach taken by Feng et al. [14] constructed a so-called “universal” object detector using a combination of concept detectors with small semantic gaps. In [21], Jiang et al. demonstrated the benefit of leveraging both high and low-level features for multimedia event detection.

In the present work, we model the events using various low-level features which are extracted from the given set of videos itself. We also include a set of object models independent of the events, which can help in capturing some semantic information. In addition, a weakly supervised semantic concept detection is employed to retrieve semantic information specific to the target events.

2.2 Evidence description for video retrieval

Along with video retrieval/classification tasks, the problem of providing a description of the important evidence that a video clip or an image contains an instance of a semantic category has been steadily gaining prominence in the computer vision community. For example, a number of papers have been proposed to leverage latent topic models on low-level features [5,8,44,55]. To date, the most common approach to such lingual description of images has been to model the joint distribution over low-level image features and language, typically nouns. Early work on multimodal topic models by Blei et al. [5] and subsequent extensions [8,15,44,55] jointly model image features (predominantly SIFT and HOG derivatives) and language words as mixed memberships over latent topics with considerable success. Other non-parametric nearest-neighbor and label transfer methods, such as Makadia et al. [38] and TagProp [17], rely on large annotated sets to generate descriptions from similar samples.

Alternatively, a second class of approaches directly seeks a set of high-level concepts, typically objects but possibly others such as scene categories. Prominent among object detectors are Object Bank (OB) [34] and the related deformable parts model (DPM) [13] which have been successful in the task of annotating natural images.

While this paper focuses on the problem of multimedia event detection, our LSVM method also additionally enables discriminative summary by temporal evidence localization for the target events. It is noted that the proposed scheme of temporal evidence localization, represented with discriminative video segments for an event class of interest, are different from those considered for the TRECVID MER tasks [42,43], which are based on lingual description of key evidence. When discriminative clusters of LSVM model are manually named, the temporal evidence localization can be further extended to provide semantic textual description as well.

2.3 Score fusion

When addressing the problem of event detection, we can benefit from exploiting different modalities including video, audio, and textual channels. In [22], Jiang and Loui extended upon a method that considered concept detection in 10-second video segments to construct an event recognition system that operates on an entire video using a global bag of "audio-visual grouplets" representation. In our work, we compute an array of features to represent visual as well as audio channels. However, unlike [22], we do not group the features together at an early stage, but adopt a late fusion scheme to combine the scores from individual base classifiers, which is called *score fusion*.

There exist numerous score fusion methods, which can be grouped into three main categories. The first category can be understood as *blind* fusion where fixed rules are applied regardless of actual base classifier score distributions, prior to simple score summation. As one of the pioneering works, multiple classifier combination rules are studied in [28], where extensive experiments showed that Sum and Product are top two best performing methods. In recent work [50], the geometric mean is reported to be highly effective despite its simplicity. Both product and geometric mean can still be understood to belong to the first category where a logarithm transformation is used prior to summation. While simplicity is the main advantage, as also reported in [50], more sophisticated fusion methods can often outperform them at the expense of additional computation.

The second category of late fusion methods [20,47,40] are formulated within a score normalization framework, where particular assumptions are made on score distributions and used to align base classifier scores. However, most of these methods require the normalization transforms to be determined manually, based on expert knowledge. Our work differs in that it is more focused on building a robust fusion model from a large set of black-box classifiers.

The third category aims to learn a score function to combine scores from multiple base classifiers. In [51], weights are learned by minimizing different target error metrics with different regularizations. In [36], a linear dependency between features is proposed to address the independent assumption issue in fusion process. Smith et al. [48] treat the confidence scores from multiple models as a feature vector, and then learn a classifier for different classes using a sample-based approach. Lan et al. [30] introduced a double fusion technique, which unifies both feature level fusion and score level fusion. Our two fusion learning algorithms [26,35] introduced in Sect. 6 belong to the third category with the following distinctive properties: MFoM method [26] can optimize learning process for a wide array of custom metrics [26]; Local Expert Forest [35] learns multiple localized fusion functions across multi-dimensional score space rather than relying on a single fusion function.

3 Overview

Our system consists of three major building blocks: (1) feature extraction, (2) base classifiers, and (3) score fusion, as illustrated in Fig. 1.

Real-world videos exhibit significant variations in salient sensory information across different event categories. For example, 'parkour' event is distinctively characterized by motion in videos, while 'birthday party' event exhibits unique auditory patterns from birthday songs.

Therefore, our feature extraction module (shown in orange in Fig. 1) computes a large array of multimodal features (both visual and audio) from input videos, which improves its capability to capture salient information across diverse event classes. The features incorporated in our system are described in Sect. 4.

Next, multiple *base classifiers* independently compute detection scores based on available features. As a result, each video clip is associated with multiple base classifier scores. In our framework, each base classifier can incorporate an arbitrary subset of the features (i.e., anywhere from one to all features), and the same features can be re-used across base classifiers. For each base classifier, its parameters are learned and tuned via cross-validation on training data.

Most of the base classifiers incorporated in our current system are based on SVMs or variations of SVMs such as Multiple Kernel SVMs or Latent SVMs (LSVM). In particular, we use LSVMs in a novel high-level scene concept detection and localization framework, which enables unique summary descriptions of retrieval results (see Sect. 7.2.2 for video description results). More details of the base classifier types and incorporated kernels are described in Sect. 5.

The major rationale for incorporating multiple base classifiers stems from two considerations: computational demand

and open expandable architecture. In terms of computational demand, simply, multiple base classifiers operating across different subsets of features is more parallelizable and less memory intensive, compared to the alternative of loading and using entire features at the same time, such as early fusion [3,23]. Furthermore, different base classifiers may use different modeling schemes, e.g., SVMs or logistic regression, and provide multiple instantiations of classifiers that are designed to be fused through the next step of score fusion results in being more open and expandable, which can incorporate additional off-the-shelf classifiers in a flexible manner as needed.

Finally, the *score fusion* module combines multiple base classifier scores through diverse fusion methods, and computes a single final detection score per video clip. We incorporate multiple fusion methods in our system, both untrained (average and geometric mean fusion) and learning-based (Local Expert Forest [35] and Maximal Figure-of-Merit [26]). The parameters of the learning-based fusion methods are estimated through cross-validation. Interestingly, our study indicates that there is no clear winner among different fusion methods. Accordingly, our system incorporates a *Mix-and-Match* framework, where different fusion approaches are tried on each event class, and the best approach per class is selected. More details of the fusion learning scheme and novel fusion methods are described in Sect. 6.

4 Features

The proposed system incorporates a large array of audio-visual features to improves its capability of capturing salient information across diverse event classes. In the following sections, incorporated visual (Sect. 4.1) and audio (Sect. 4.2) features are described. For each modality, features are categorized to be either “low-level” or “mid-level”. In this work, “high-level” features are exclusively utilized by Latent SVM in Sect. 5.2.

By low-level features, we mean features such as Color SIFT (CSIFT) [46] which do not directly deliver semantic information on its own. On the other hand, mid-level features directly detect semantic categories such as visual object classes (e.g., people and vehicles) or semantic auditory patterns (e.g., drilling sound or laughter).

In this paper, most low-level features are used in the form of standard bag-of-words (BoW). That is, the raw visual and audio features are first computed over small spatial/temporal volumes densely sampled across a given video clip. A large, unlabeled collection of these features are clustered to build a codebook, and each computed feature is quantized by the cluster index in this codebook. Finally, the quantized features are pooled spatially over each frame and/or temporally across the entire clip, producing a summary BoW feature vector.

In this work, most of the features are pooled into a single feature vector for the entire clip. The exception is for our LSVM model: we pool the features across small temporal segments to allow for temporal localization (see Sect. 5.2).

4.1 Visual features

4.1.1 Low-level features

The list of our visual features includes HoG3D [29], GIST [41], Color SIFT [46], independent subspace analysis (ISA) [31], transformed color histogram (TCH) [46], and a set of visual features from [57] (which we call “SUN09” in this work), including histogram of gradients (HoG), geometry texton histogram (GTH), self-similarity measure, dense/sparse SIFT, local binary patterns (LBP), and tiny image.

Each low-level feature is introduced to capture different types of visual information. Their properties are summarized in Table 1. For example, ISA feature is capable of capturing temporal dynamics in video sequences [M], constructed as a BoW vector [B], learned in an unsupervised manner in a training corpus [U], and designed to describe an entire keyframe image-scene [S]. We have observed that this large set of low-level features is helpful to detect generic multimedia event classes, while they are mostly complementary in our fusion methods, discussed in Sect. 6. The performance of individual features can be found in Sect. 7.

4.1.2 Mid-level concept features: Object Bank

To directly capture visual semantic information from videos, our system explore the use of a large set of visual object detectors, which are provided by Object Bank (OB) [34]. The integrated implementation utilizes a set of visual object detectors based on histogram-of-gradients (HoG) templates [10], which are scanned across uniformly sampled video frames, and agglomerated detector responses across each video clip are used to categorize events. The overall computational process of OB is shown in Fig. 2. An array of object detectors are first applied to each image at multiple scales. The filter responses from each scale are represented as a three-level spatial pyramid and are accumulated to produce high-dimensional scene content descriptors per image.

In our work, object detections are regarded as *mid-level concepts*, in the sense that the object categories can be shared across complex events. Accordingly, OB utilizes the distribution of mid-level concepts to characterize events, and even “high-level” concepts are built upon OB features as described later in Sect. 5.2.

Compared to traditional scene-level concepts such as LSCOM [1], OB features provide semantic and descriptive understanding of visual scenes at *object level*.

Table 1 Properties of low-level visual features

	Property							
Feature	C	T	G	M	B	U	P	S
HoG3D	·	·	x	x	x	·	x	·
GIST	·	x	·	·	·	·	·	x
Color SIFT	x	·	x	·	x	·	x	·
ISA	·	·	·	x	x	x	·	x
TCH	x	·	·	·	x	·	x	·
HoG	·	·	x	·	x	·	x	·
GTH[a]	·	x	·	·	x	·	x	·
Self-similarity[a]	·	·	·	·	·	·	x	·
Dense SIFT[a]	·	·	x	·	x	·	x	·
Sparse SIFT[a]	·	·	x	·	x	·	x	·
LBP[a]	·	x	·	·	x	·	x	·
Tiny image[a]	x	·	·	·	·	·	·	x

C color, *T* texture, *G* gradient, *M* temporal, *B* BoW, *U* unsupervised learning, *P* patch-based, *S* entire image scene

[a] SUN09 features

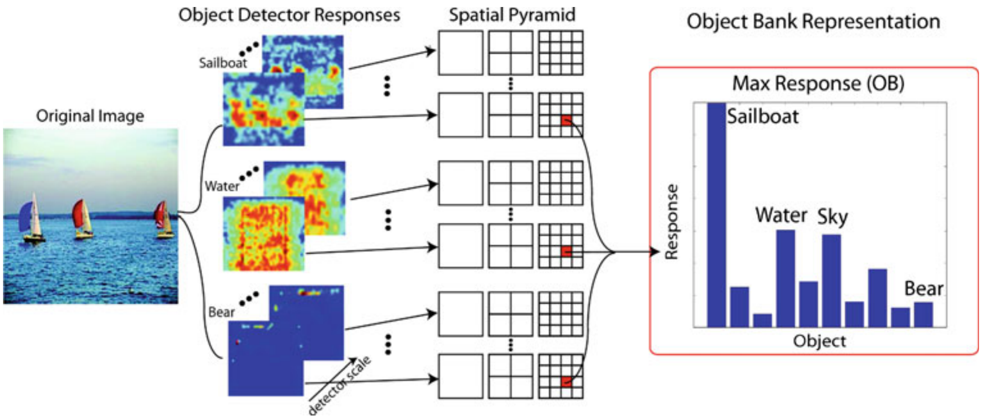


Fig. 2 Illustration of Object Bank (image courtesy of [34]). Original Object Bank feature is computed by accumulating multiple semantic object detector responses across multiple scales and spatial layouts

The current OB implementation incorporates detectors for 177 object classes and is one of the main large-scale object recognition system publicly available. These object classes are independent of the event categories used in our experiments in Sect. 7.

Our system system discards spatial pyramid layout information, because we find that the variation of object locations within unconstrained videos is not regularized and resulting lower-dimensional representation helps the generalization capability during classifier training. Due to the large number of scanning windows across multiple scales, the computational demand of OB features is very high compared to the low-level features. Accordingly, OB features are applied on temporally uniformly sampled frames, e.g., frames at every 2 s.

Our work explored various representations of OB features. For example, to provide features for standard SVM classifiers (see Sect. 5.1), OB features across frames are pooled to build clip-level descriptors. Two variations of pooling are used in our work: maximum and average pooling. Furthermore, frame-wise OB features are used for weakly supervised temporal Latent SVM model presented in Sect. 5.2.

4.2 Audio features

4.2.1 Low-level feature: MFCCs

To capture general audio information of a video, we use MFCCs with a BoW model. In particular, 32-dimensional MFCCs are extracted at every 10 ms with 25 ms frame size. Then, MFCC features are quantized based on a codebook with 1 K size using hard assignment.

4.2.2 Mid-level feature: acoustic segment models

A conventional way of exploiting audio semantics for MED is to use a set of pre-defined audio concepts [33,7]. However, using a fixed set of audio concepts to perform event detection might not be suitable because consumer-level videos tend to be unconstrained and unstructured. As such, there exists a wider range of variability in audio signals.

We developed acoustic segment models (ASMs) [6] to understand a broader range of mid-level audio semantics by capturing the diverse temporal structures within low-level audio signals. ASMs build upon previous work such as fundamental speech sound units for speech recognition [32], which have been applied to music genre classification [45] and speaker recognition [52]. Our approach is the first study of ASMs to MED by building bottom-up acoustic semantic words. In particular, unlike previous work that exploits temporal acoustic structure in particular domains, e.g., speech or music, the developed ASMs provide an extended framework for generic audio sound types.

We modeled ASMs as 3-state HMMs. They were trained with a set of ‘representative’ audio segments for given multimedia event classes; in particular, 8 initial segments were manually chosen from an event class. For example, initial segments for the ‘Birthday Party’ event class include singing a birthday song, cheering, laughing, and clapping, while those for the ‘Getting a Vehicle Unstuck’ event class include tire spinning, motor, and street noise. Then, we iteratively conducted Viterbi decoding and Baum–Welch estimation to refine the models until they converged. The typical length of decoded segments is 100–200 ms. Once ASMs are obtained, each audio clip is transformed into a BoW vector by considering N-best Viterbi sequences with unigram and bigram statistics. More details of modeling and learning of our ASMs can be found in [6].

5 Base classifiers

Once features are extracted, multiple base classifiers produce scores independently, based on different model formulations and different subsets of features.

Concretely, it is assumed that, for each event category, a training set of N videos, $\{(x_i, y_i)\}_{i\in 1\ldots N}$ is available, where x_i is the ith video and $y_i \in \{-1, 1\}$ is its label. For brevity, let’s reuse x_i to denote the entire set of computed features for the ith clip, where subscripts will be omitted if not necessary, i.e., x.

Additionally, let’s denote the entire set of base classifier functions as $\{f_k\}_{k\in 1\ldots|k|}$ where f_k is each base classifier. Finally, each base classifier f_k computes a classification score z_k based on a corresponding subset of features $x_k \subset x$ as follows

$$z_k = f_k(x_k) \quad \text{where} \quad x_k \subset x \tag{1}$$

Through the remainder of the draft, the entire set of base classifier will be denoted by $Z_{\text{B}} = \{z_k\}$.

Each base classifier in our system is learned in an one-vs-all manner for the target event based on available query exemplar videos and a common set of background videos that are used as negative training data. As mentioned in Sect. 3, the computed set of multiple base classifier scores Z_{B} will be later combined to a single detection score Z_{F} via score fusion (Sect. 6).

We learn three different types of base classifiers: (1) support vector machines (SVMs) [9] using linear or non-linear kernels (e.g., histogram intersection kernel), (2) multiple kernel learning (MKL) [53] to learn a classifier across multiple features jointly, and (3) a temporal variation of Latent SVM model. The details of each approach are described in the following sections.

5.1 SVMs and multiple kernel learning

One of the advantages of using SVMs as base classifiers is that decision boundaries in feature space can be learned within a max-margin framework, which potentially improves generalization power during the training process. In addition to linear SVMs, we use three types of kernels across for non-linear SVMs: the histogram intersection kernel (HIK) and the χ^2 kernel (Chi2), which are commonly used in computer vision, and the negative geodesic distance kernel (NGD) [60], which empirically shows competitive performance for histogram-based features by assuming the features have multinomial distributions. We apply different kernels even on identical features, and we observe non-trivial differences in performance (see Sect. 7 for details).

In addition to standard SVMs on single features, our system incorporates MKL SVMs to combine the discriminative power across a set of visual features. Different from the score fusion framework discussed in Sect. 3, MKL combines kernels computed across features into a single kernel value through weighted summation. In particular, we use the Simple MKL implementation from [53], which learns weights to be used during kernel combinations. The training of a MKL SVM is computationally demanding, so we heuristically select a set of features for MKL based on prior work [57]; these are marked as SUN09 features in Table 1. For MKL, we used Chi2 kernels.

Both for non-linear SVMs and MKL, while they demonstrate a clear performance advantage over linear SVMs, the run-time computational demand for non-linear models is very high, both for training and testing. This is especially so when dataset is large, features are high-dimensional, and there are a large number of support vectors. To address the heavy computational demand required to use non-linear SVMs, our system incorporates two recently introduced approximation techniques [37,54]. In particular, these techniques are applicable for additive kernels such as HIK and Chi2, and improves the speed during both training and testing phases significantly. First, explicit feature maps [54] for the additive kernels are used. This technique expands original features into a higher-dimensional space, but allows the much more efficient linear SVM to be used during training and testing, with minor approximation error, which results in a negligible performance drop compared to using exact kernels with non-linear SVM training. In addition, we incorporate [37], which builds lookup table using piecewise approximation, effectively providing constant time execution during testing, regardless of the number of support vectors.

5.2 Latent SVM

Our temporal variant of Latent SVM is a new development which learns salient parts of videos and detects "important" temporal regions in test clips, resulting in improvement in performance and capabilities of discriminative summary descriptions for retrieval results. The overall formulation is analogous to the original formulation of spatial Latent SVM model for object detection [13]. In this paper, the framework has been used with Object Bank (OB) feature only, since the model utilizes "high-level" concepts which are built in an unsupervised manner, bottom-up from mid-level concepts such as OB. Nonetheless, the framework is general, and is not limited to OB features.

5.2.1 High-level visual scene concept learning

The idea here is to build upon the base mid-level features in a hierarchical fashion to represent an event as a series of high-level visual scene concepts. These concepts are typically class dependent and require trained models for each concept in every class. Under a traditional supervised learning approach, segment-level annotations would be required to train such models. Instead, we used an unsupervised technique to extract class specific concepts.

In particular, OB features are extracted for a dense set of frames, e.g., sampled at every two seconds. Then, clustering all frame-level OB features from an event class produces clusters of video frames with similar scene objects. In other words, clusters represent different visual "scene type" concepts for a class. Concretely, scenes involving snow, sand, mud, or flooded streets may form distinct high-level concept clusters for the "getting a vehicle unstuck" event.

After extracting event-specific scene clusters, a classifier can be trained for each scene concept. For this purpose, we use all segments from the training videos and assign them to the closest scene type cluster using Euclidean distance. For each cluster, an HIK-SVM classifier is trained to separate one scene type from all other clusters, as well as negative video frames from other event categories using a bootstrapping approach.

In addition to such scene concepts, a clip-level feature vector is also used to capture the overall context of a video, where OB features are agglomerated across all frames by average pooling.

5.2.2 Classification using latent SVMs

The high-level scene concepts described in Sect. 5.2.1 can be leveraged for MED. When designing such an MED system, a first observation is that only a subset of scene concepts will appear in any given video. For example, a user-uploaded video containing a vehicle being unstuck may show an environment filled with snow, mud, or sand, but it will seldom include all three. Accordingly, a video will be represented as a small subset of its most discriminative mid-level concepts.

An event of interest typically occurs only in a small sub-section of a sequence. Even though the entire temporal domain of a video provides background context for an event, considering only this global domain in isolation can be misleading due to overwhelming amounts of unrelated content (e.g., the actual act of freeing a vehicle in the "getting a vehicle unstuck" class may occupy a few seconds in a much longer traffic-related sequence). Furthermore, due to the large diversity in multimedia event classes, enforcing a fixed temporal order of smaller sub-activities may prove to be too rigid. Accordingly, we ignore the temporal ordering of scene concepts and focus on determining if sufficient evidence can be localized that is indicative of a specific event class.

In detail, the goal is to learn a scoring function $f(x)$ in in Eq. 1. Note that subscript k is omitted for brevity. Concretely, the proposed model f with model parameters $w = \{w_g, \bigcup_{s=1}^{s=S} w_s\}$ consists of two terms as follows:

$$f(x, h|w) = w_g^T \phi_g(x) + \sum_{s=1}^{S} b_s w_s^T \phi_s(x, t_s), \tag{2}$$

The first term is a global model that captures the total theme of the video. The second term describes a local model that represents an event by a set of scene concepts. Together, these two terms form the full model, where $\phi_g(x)$ is a global feature extracted for sequence x, and w_g is the corresponding weighting vector. Additionally, $\phi_s(x, t_s)$ and w_s denote the feature and weight vectors for a particular scene concept s. As discussed in Sect. 5.2.1, concept detectors have been trained that provide a scene label score $A_s(x, t_s)$, for a particular frame of the video x, centered at time t_s. We use this compact scene label score for a scene concept feature vector with a bias term, i.e., $\phi_s(x, t_s) = [A_s(x, t_s) \;\; 1]$. Finally, in Eq. 2, b_s is a binary variable that indicates whether a scene concept type is present in the video. The proposed Latent SVM framework

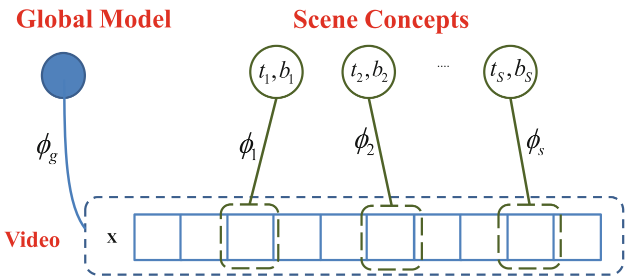


Fig. 3 Depiction of the proposed Latent SVM framework. The global model (ϕ_g) captures the overall context of a video, and the scene models ($\phi_1, \ldots, \phi_s$) represents the different scene types observed in the category. The presence of a scene type is represented using binary variables b_s, and the temporal position of scene types in a video is denoted by t_s

with this global and scene concept features is illustrated in Fig. 3.

In the proposed model in Eq. 2, the b_s and t_s variables allow an event to be represented as a subset of scene concepts that can be localized in the sequence. These variables are *not* provided at all during training and testing, hence, they are hidden *latent* variables as $h = \{(b_i, t_i)\}_{i \in 1 \ldots S}$. In particular, during testing, the localized concepts (b_s) and their locations (t_s) effectively provide *discriminative descriptions* for each retrieval. Thus, the final scoring function is

$$f(x|w) = \text{argmax}_h f(x, h|w)$$
$$\text{s.t.} \quad \sum_{s=1}^{S} b_s = M \tag{3}$$

where the constraint is added to ensure that only M scene concepts are used to represent the sequence (recall it is assumed that only a small subset of scene concepts are present in any sequence). The maximization in Eq. 3 can be performed in two steps. First, all values of t are enumerated and the one that maximizes $w_s^T \phi_s(x, t_s)$ is chosen for each scene type, s. Then, the M scene types with the highest score are selected, and only their corresponding binary variables, b_s are assigned a value of 1.

Finally, the parameters of the model are trained in the Latent SVM [13,56] framework:

$$\min_{w,b} \frac{\lambda}{2} ||w||^2 + \sum_i^n \xi_i$$
$$\text{s.t.} \quad y_i(F_w(x_i) + b) > 1 - \xi_i, \tag{4}$$

where ξ_i are the standard SVM slack variables.

6 Fusion

With a set of base classifier scores in hand, i.e., $Z_B = \{z_k\}$ from Eq. 1, the final step in our system is score fusion, where these scores are combined into a single score Z_F for the clip. Concretely, this process is conducted by a fusion function F:

$$Z_F = F(\{z_k\}) \quad \text{where} \quad z_k = f_k(x) \tag{5}$$

We developed a novel scheme to facilitate score fusion learning, for when the amount of training data is limited. A traditional approach for fusion learning is to divide available data separately for base classifier training and fusion learning, which results in less data for both learning tasks. In comparison, our scheme allows us to maximally utilize every training exemplar during both training phases, as described in Sect. 6.1.

For score fusion and learning thereof, we developed two novel discriminative fusion learning algorithms: Maximal Figure-of-Merit (MFoM) [26], and Local Expert Forest (LEF) [35] learning schemes, which will be discussed in Sects. 6.2 and 6.3.

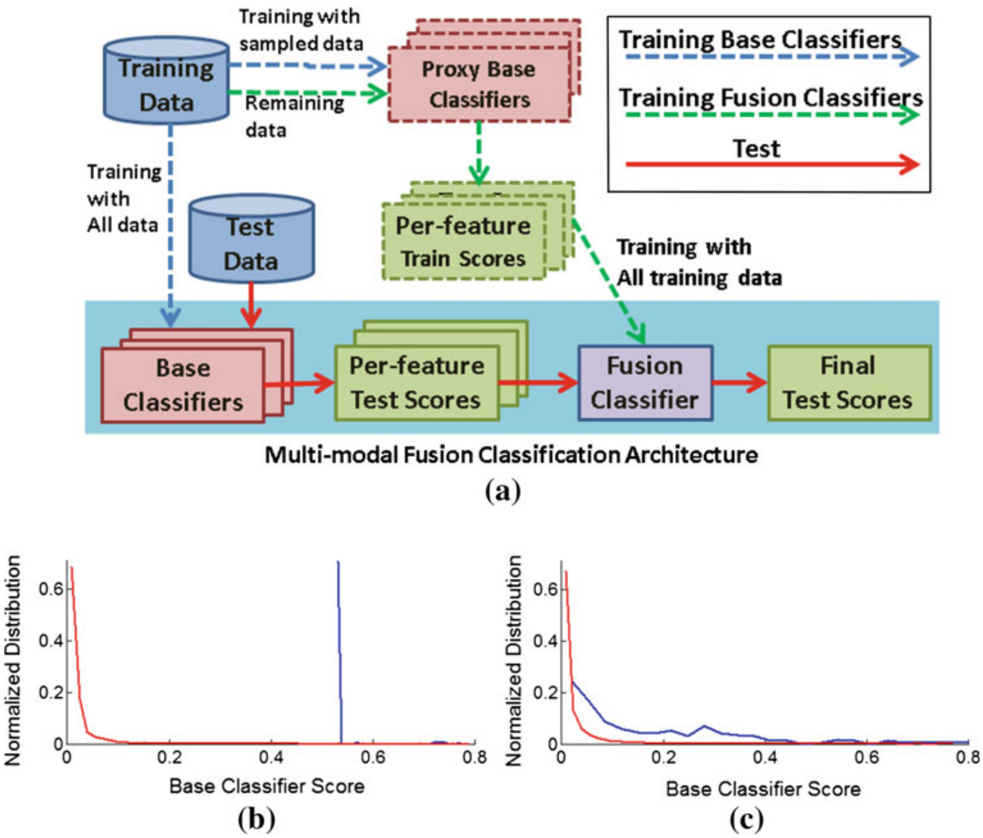
In addition, our study indicates that there is no clear winner among different fusion methods. Accordingly, our system incorporates a *Mix-and-Match* framework, which is described in Sect. 6.4.

6.1 A robust fusion learning scheme for limited training data condition

The overall framework for our robust fusion learning framework is illustrated in Fig. 4a, where three separate data flows are shown: "proxy" base classifier training (blue dashed), fusion classifier training (green dashed), and test phase (solid red), respectively. During the test phase, the classification system at the bottom in Fig. 4a applies base classifiers on test data to produce per-feature test scores, e.g., audio and video, independently. These base classifiers are trained a priori using all available training data. Then, these scores are concatenated and used as an input vector to a fusion classifier which produces a single final score.

Under our robust fusion learning scheme, our system divides training data into sets where they are used separately to train proxy base classifiers and a fusion classifier, which is designed to maximally use available training data for fusion learning. By proxy base classifiers, we mean temporarily constructed base classifiers which are learned from a subset of available training data. Subsequently, remaining training data (not used for proxy training) are fed into these proxy base classifiers and scores are generated, which are then used as training inputs to learn fusion classifiers.

In detail, it is tempting to apply the "full" base classifiers shown at the bottom left of Fig. 4a on the already used training data to produce base classifier outputs to be used as fusion classifier training data. However, this approach fails to learn an accurate fusion classifier. The reason is that the base classifier has already seen all the training data, accordingly, the



(a)

(b)

(c)

Fig. 4 **a** The proposed robust fusion framework, with separate data flows for training and test phases. Comparison between score distributions from a base classifier on **b** training data seen during learning the base classifier and **c** unseen test data; *Blue and red lines* indicate distributions of positive and negative samples, respectively. There exists inconsistency between scores in **b** and **c**; scores in (**b**) are unrealistically accurate, and not suitable to be used to train a fusion classifier (color figure online)

generated outputs are unrealistically accurate especially with non-linear kernels such as histogram intersection or negative geodesic distance kernels. For example, Fig. 4b, c show score distributions generated by a base classifier on already seen training data and unseen test data, respectively. In particular, Fig. 4b shows more realistic spread-out score distribution on unseen test data. Because these two distributions are distinct, a fusion classifier learned from the unrealistically accurate scores shown in Fig. 4b is unlikely to perform well on novel data.

Our solution is illustrated in Fig. 4a as dashed training flows. In detail, training data is divided into N subsets and proxy base classifiers are learned with $(N-1)$ subsets, then used to generate scores on a remaining subset. This procedure is repeated N times to generate scores for entire training data. This way, we can obtain more realistic base classifier outputs to be used to train a fusion classifier.

6.2 Maximal figure of merit (MFoM)

In real-world retrieval tasks, the performance metrics that capture user desires can differ widely from application to application. However, a conventional approach to learning such domain-specific performance metrics is to blindly minimize standard error rates and hope the targeted metrics improve, which is clearly sub-optimal. To address this issue, we use a fusion framework based on MFoM learning [16], which is able to directly optimize specific performance metrics.

In this work, we focus on the weighted sum of the probabilities of missed detections (P_{MD}) and false alarms (P_{FA}) at a particular ratio, suggested by the TRECVID '11 MED task [2]. Concretely, the goal is:

$$\text{Minimize } S_\tau = P_{\mathrm{MD}} + \tau \times P_{\mathrm{FA}} \text{ s.t. } \frac{P_{\mathrm{MD}}}{P_{\mathrm{FA}}} = \tau, \tag{6}$$

where τ is a desired ratio of P_{MD} and P_{FA}. However, it is noted that the MFoM learning method is general beyond this particular metric and easily applied to other popular metrics such as F_1-score and average precision [16,25].

In particular, we adopted a linear fusion model, which is widely used for late fusion frameworks, due to its simple and straightforward manner to analyzing confidence and correlation of base-classifier scores. A final fusion score of a sample

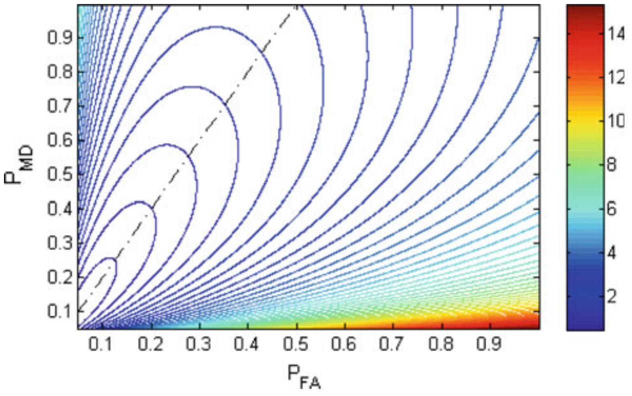


Fig. 5 Iso-contour curves of the loss function $L(T;\Lambda)$ that simulates the target metric in Eq. 6. The *dashed straight line* corresponds to a iso-ratio $\widehat{P_{\mathrm{MD}}}/\widehat{P_{\mathrm{FA}}} = \tau$, when $\tau = 2$. *Left bottom corner* corresponds to perfect classification

x given a set of model parameters Λ is formulated in a linear discriminant function (LDF) format as

$$Z_{\mathrm{F}} = F(\{z_k\}|\omega) = \sum_{k=1}^{|k|} \omega_k z_k + \omega_0, \tag{7}$$

where z_k indicates a score of x from the k-th base classifier, and ω_k and ω_0 are a corresponding weight for the base-classifier score and a bias term, respectively. Accordingly, our goal is systematically learning the linear weights for each score dimension.

The core ideas of our MFoM-based learning approach are twofold. First, we exploit the fact that most custom target metrics and their sub-components, such as P_{MD} and P_{FA} in Eq. 6, can be expressed as a combination of the four sub-metrics from a confusion matrix, i.e., TP, FP, TN, and FN. Second, we approximate a target metric such as S_τ in Eq. 6, that is based on discrete error counts—making it challenging to apply advanced optimization techniques—with a parameterized continuous and differentiable loss function $L(T;\Lambda)$, where $T = \{(x_i, y_i)\}$ denotes a training corpus, and Λ is a set of parameters for approximation and ω (refer to [26] for details). Then, the optimal parameter Λ_{opt} that minimizes $L(T;\Lambda)$ is learned by gradient descent algorithms such as the generalized probabilistic descent (GPD) [24], where the GPD conducts iterative descent steps with varying learning rate. The details of designing the loss function $L(T;\Lambda)$ that approximates the target metric in Eq. 6 in the MFoM framework are presented in [27].

To showcase the quality of the approximated loss function $L(T;\Lambda)$, Fig. 5 illustrates the iso-contour curves of the loss function, along with the dashed line which corresponds to the ratio constraint for the sample case of $\tau = 2$. It can be clearly seen that the designed loss function is correlated with and declines towards the iso-ratio line. This implies that the learning process will driven towards the minimum value near the iso-ratio line and left-bottom of the plot as desired.

In our previous work [27], we have observed that this MFoM-based score fusion outperforms other linear fusion methods including logistic regression and linear-SVM fusion, which do not optimize a domain-specific metric, by about 7.3–12.9 % relatively, on average.

6.3 Local expert forest

Many score fusion techniques (e.g. [51]) including our MFoM approach [26] (Sect. 6.2), build global linear models where each base classifier's output is globally weighted. For learned models such as MFoM, the weight assigned to each classifier's output is proportional to its performance on training data, and untrained fusion by geometric or arithmetic mean use uniform weights. These methods perform quite well in many cases, but do not account for local performance variations. If classifier 1 is assigned weight w_1, and classifier 2 is assigned weight $w_2 > w_1$, the model asserts that classifier 2 is more reliable *regardless of the score value*. Figure 6 illustrates a case where this does not make sense because there is no consistent ordering of classifier performance. While both ObjectBank and HoG3D outperform the GIST base classifier in this case, the relative performance of the former two is more complicated. For clips where both these base classifiers give scores >0.5 (i.e., the upper-left part of the DET curve), ObjectBank outperforms HoG3D. However, when both give scores <0.5, HoG3D outperforms ObjectBank. Whereas a global fusion model would learn equal weighting for the two, we have developed a model that learns such local performance variations, giving a higher weight to classifiers that perform better in a local region of the score space. For a system with N base classifiers, each clip is considered to be a point in an N-dimensional score space. Within local regions of this score space, we learn sets of weights that reflect the local performance of the base classifiers. In the style of random forests, we find several such weight sets for different pseudo-random partitions of the score space, and average the results from each local weighting at test time. A graphical depiction of the model is shown in Fig. 7, and additional details are presented in [35]. Like the general mixture of expert model, ours is formulated as

$$Z_{\mathrm{F}} = F(Z_{\mathrm{B}}) = \sum_{E} P(Y = 1|E, Z_{\mathrm{B}})P(E|Z_{\mathrm{B}}), \tag{8}$$

where Y is the unknown label of the data and $P(E|Z_{\mathrm{B}})$ is the 'gate' function, indicating which sub-model is responsible for generating each data. We use linear models for local experts, with a likelihood function from the ith expert with a parameter set $w^{(i)}$ being

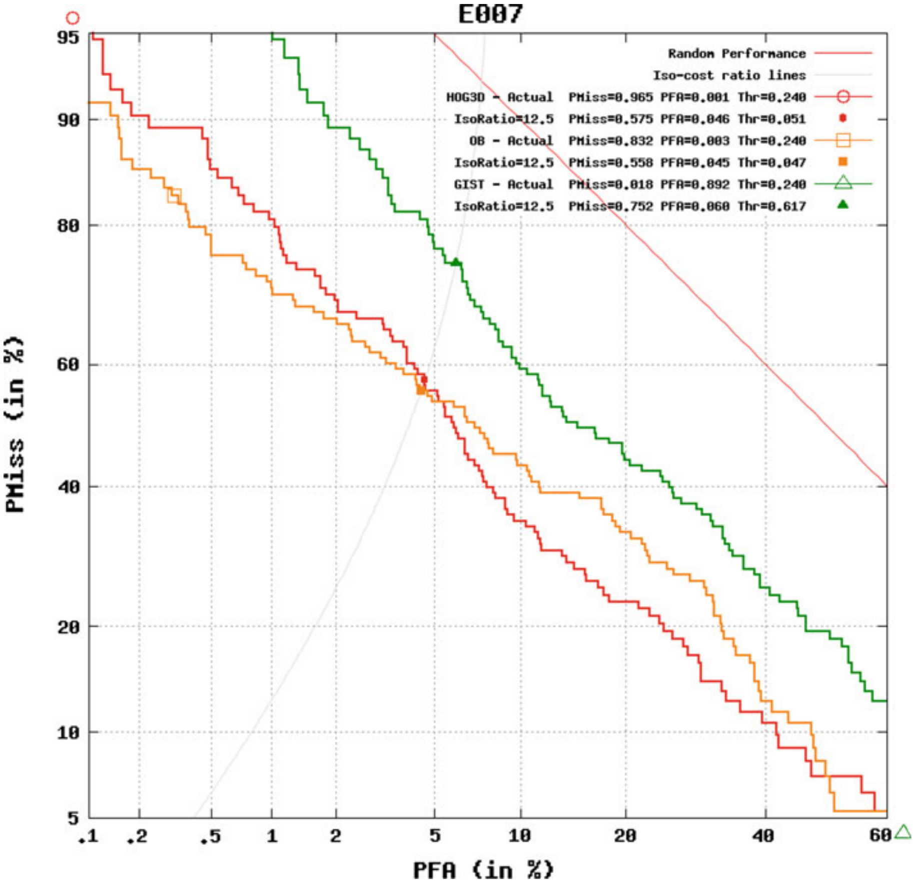


Fig. 6 Motivation for a local fusion model. While certain base classifiers consistently outperform others, e.g. HoG3D (*red*) and ObjectBank (*orange*) beating GIST (*green*), the performance ranks of others may vary as a function of score. In cases where DET curves cross, as HoG3D and ObjectBank do here, we learn a fusion model that accounts for changes in performance rank. Each *curve* shows performance for different target event where the *x*-axis corresponds to probability of false alarm, and *y*-axis indicates probability of miss, i.e., *lower left corner* corresponds to perfect retrieval (color figure online)

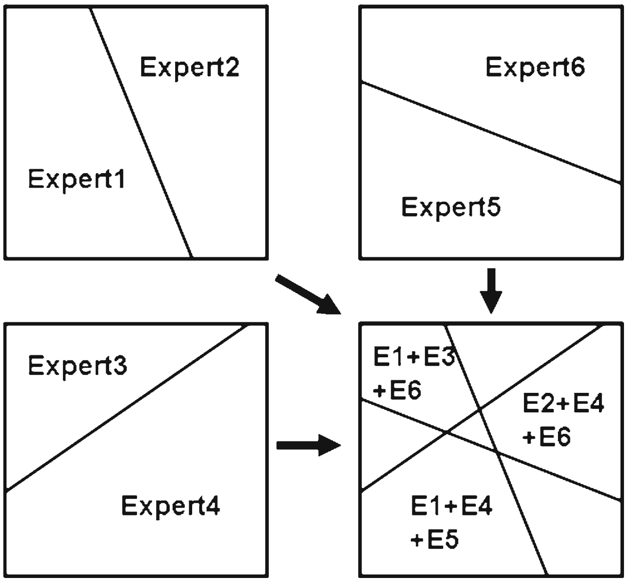


Fig. 7 Illustration of local expert forest-based fusion. For a notional 2D score space (i.e., fusion of two base classifiers), we generate multiple binary partitions and learn different sets of linear weights for test data within each partition (*black boxes*). At test time, the multiple weighted sums from each binary partition are averaged to generate the fused score

$$P(Y|E^{(i)}, Z_{\mathrm{B}}, w^{(i)}) = \exp\left\{-\|Z_{\mathrm{B}} \cdot w^{(i)} - Y\|^2\right\}. \tag{9}$$

During learning, the parameters of the LEF model are optimized by a regularized minimization of Minimum Mean Squared Error described in [35].

6.4 Mixing and matching fusion models

The various score fusion models that we have developed, and the standard methods that we have considered, have different strengths. The untrained fusion methods, such as geometric and arithmetic mean, are a good choice when the semantic event class is very diverse and only partially spanned by the available training data. The MFoM model, due to its explicit target-metric optimization, provides competitive performance with a simple LDF framework for a targeted metric. The Expert Forest model provides improved performance when base classifiers lack a consistent performance ordering, but may over-fit to limited training data due to the increased model complexity.

In order to provide the best system performance for a wide range of events, we apply a ‘mix-and-match’ approach to select the best fusion model for each event. Using the validation method described above, we train multiple fusion models for each event, and evaluate these (and the untrained fusion methods) against held-out training data. We choose the fusion model which provides the best fused performance. For the untrained fusion methods, we further evaluate the fusion performance over subsets of the available base classifiers, and choose the method that performs the best for each event. In contrast, because of the learned fusion model's ability to implicitly perform classifier selection (by assigning a weight of zero), we do not explicitly search for the best base classifier inputs for the trained models.

7 Experimental results and analysis

7.1 Data and metrics

For evaluation of our approaches, we have conducted experiments on the TRECVID Multimedia Event Detection (MED) 2011 test set [42], which includes labeled instances of 10 complex event classes in an archive of video clips representing over 1,000 h of video. The list of the 10 event classes are shown in Table 2. This is one of the largest dataset of unconstrained videos for which ground truth is available.[2] For each event class, there are approximately 100 positive training exemplars, along with about 8,000 negative training dataset which does not belong to any class. The size of the test archive is about 32 K video clips, with roughly 100 positive test clips per class. Note that positive clips for each class only constitute 0.37 % of the test clips, with the remaining majority of the clips constituting negative background dataset, providing a realistic proxy for real-world retrieval problems.

Table 2 List of MED'11 events

IDs	Event class
E006	Birthday party
E007	Changing a vehicle tire
E008	Flash mob
E009	Getting a vehicle unstuck
E010	Grooming an animal
E011	Making a sandwich
E012	Parade
E013	Parkour
E014	Repairing an appliance
E015	Sewing project

In line with the TRECVID evaluation rules, our system observes event independence. That is, positive training examples are only used to learn base classifier and fusion models for the event class to which it belongs. Moreover, with respect to fusion, we generate fused scores for a given event class using only base classifiers trained to detect that class. When fusing scores to generate a final clip score for 'birthday party', for instance, we ignore scores for 'repairing an appliance', despite the prior knowledge that clips are unlikely to depict both activities. However, we have previously shown that incorporating non-target event scores in fusion improves performance [27].

Per the TRECVID conventions, we present our results as Detection Error Tradeoff (DET) curves for the reasons outlined in [2]. In DET curves, x-axis represents false alarm ratio (FAR), and y-axis indicates the probability of missed detection (PMD) which is equal to 1—recall. Similar to ROC curves, DETs quantify system performance over a range of operating points. For DET curves, the left bottom corner represents the ideal system performance which detects all true positives while no false alarm occurs.

While DET curves provide an understanding of how the system performs, comparing DETs (e.g., between alternative fusion techniques) can be complicated by the example DET crossings illustrated in Fig. 6. It is therefore necessary to reduce DETs to a scalar metric which can be compared to choose between different approaches. At a high level, there are two approaches. The first choice is to measure system performance over a range of operating points, e.g. by computing the area under the DET curve (AUC). The problem with an AUC-type measurement is that it may be unduely impacted by system performance at the extremes of the DET, where users are unlikely to choose operating points. The second option is to measure system performance at a certain point, for instance the point at which the false and missed detections are in a given ratio. The potential disadvantage of measuring system performance at a single operating point is that the DET curve might not be smooth, and performance at a particular operating point may not be reflective of performance over even a small range of thresholds. In practice, because we find that the DET curves are smooth and that their tails have a lot of nuisance variation, we use a point-based measurement for our scalar performance measure. In the spirit of the Normalized Detection Cost (NDC) [2,42], the PMD at the point where PMD:FAR=12.5:1 is reported for our evaluation.

7.2 Base classifier results

Table 3 shows the base classifier performance over the 10 MED11 events measured by the probability of missed detection as 12.5 iso-line, where lower values indicate superior performance. For each base classifier, the features and kernel used are shown. While most base classifiers are based

[2] TRECVID MED'12 dataset is larger; however, the ground truth will not be publicly released for several years.

Table 3 Base classifier and fusion performance, measured by the probability of missed detection as 12.5 iso-line, over the 10 MED11 events (lower is better)

	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015
Base classifiers										
GIST HIK	0.667	0.685	0.326	0.558	0.736	0.593	0.524	0.544	0.551	0.691
HoG3D linear	0.645	0.775	0.341	0.484	0.724	0.700	0.452	0.519	0.590	0.728
HoG3D HIK	0.457	0.550	0.273	0.463	0.586	0.500	0.448	**0.317**	0.359	0.531
HoG3D NGD	0.462	0.468	0.258	0.474	0.529	0.479	**0.352**	0.327	0.321	0.531
ISA HIK	0.516	0.514	0.288	0.453	0.540	0.536	0.424	**0.317**	0.333	**0.506**
CSIFT linear	0.704	0.649	0.364	0.453	0.747	0.679	0.463	0.683	0.526	0.654
CSIFT HIK	0.468	0.414	**0.197**	0.368	0.540	0.493	0.359	0.413	0.372	**0.506**
CSIFT NGD	0.495	0.405	0.205	0.379	0.540	0.479	0.368	0.423	0.359	0.519
TCH linear	0.774	0.676	0.447	0.474	0.805	0.793	0.489	0.712	0.564	0.753
TCH HIK	0.532	0.532	0.273	0.389	0.575	0.543	0.398	0.404	0.397	0.593
TCH NGD	0.532	0.477	0.250	0.400	0.621	0.536	0.420	0.404	0.423	0.580
OB AVG, linear (L0)	0.645	0.550	0.311	0.432	0.644	0.600	0.519	0.625	0.590	0.617
OB MAX, linear (L0)	0.597	0.541	0.379	0.442	0.632	0.500	0.554	0.558	0.474	0.630
OB, LSVM (L1)	0.570	0.577	0.280	0.568	0.621	0.586	0.442	0.538	0.500	0.605
OB Avg, HIK	0.532	0.505	0.250	0.411	0.575	0.550	0.442	0.375	0.436	0.568
OB Max, HIK	0.516	0.477	0.250	**0.337**	0.529	**0.457**	0.429	0.442	0.385	0.519
SUN09 MKL	0.441	**0.351**	0.205	**0.337**	**0.483**	0.507	0.355	0.337	0.321	**0.506**
MFCCs linear	0.548	0.782	0.545	0.681	0.814	0.761	0.645	0.709	0.346	0.667
MFCCs HIK	0.446	0.618	0.424	0.564	0.686	0.739	0.584	0.583	0.295	0.654
MFCCs NGD	0.462	0.673	0.409	0.500	0.698	0.696	0.567	0.631	0.372	0.628
ASM186 HIK	**0.422**	0.561	0.470	0.553	0.744	0.618	0.607	0.602	**0.289**	0.553
ASM64 HIK	0.438	0.607	0.523	0.628	0.733	0.669	0.620	0.670	0.303	0.618
Fusion model										
Average	**0.265**	0.318	0.212	**0.234**	0.430	0.426	0.298	0.262	**0.184**	**0.447**
GeoMean	0.292	**0.290**	**0.189**	0.266	0.430	**0.404**	**0.281**	0.252	0.224	0.461
MFoM	0.324	0.299	0.197	0.287	0.430	0.419	0.329	0.262	0.237	0.487
LEF	**0.265**	0.318	0.197	0.245	**0.384**	0.412	0.285	**0.233**	0.211	0.461
Best base–best fusion	0.157	0.061	0.008	0.103	0.099	0.053	0.071	0.084	0.105	0.059

The best-performing base classifier and fusion model for each event are marked in bold

on a single feature, the SUN09 MKL classifier incorporates multiple features from [57], which are fused by MKL SVM using Chi2 kernel.

The results on base classifiers illustrate several important points. First, that the task really is *Multimedia* Event Detection, in that both audio and visual features are important; ASMs (an audio feature) are the top performer on 'birthday party' and 'repairing an appliance', while the other events' best performer is a video feature. Second, the use of multiple kernels is worth the computational effort; while the HoG3D HIK base classifier produces the best result on 'parkour', the NGD kernel is best on 'parade'. For most features, simply using different kernels results in non-trivial gap in performance, even for identical features. For most cases, non-linear SVMs provide significant improvement in performance over linear SVMs. However, except GIST which turns out to be fairly weak, it seems that there is no winner base classifier, which highlights the usefulness of including diverse base classifiers to deal with diverse event classes. For example, even a SUN09 MKL base classifier which combines a large subset of features enlisted in Table 1 is frequently outperformed by other single-feature base classifiers.

From the base classifier results in Table 3, we can also draw interesting observations between low-level and higher-level features. In particular, base classifiers based on even single low-level features deliver surprisingly strong performance. Overall, while higher-level features such as OB and ASMs do showcase strong performance as well, there does not seem to be any systematic advantage of either method, in terms of quantitative performance.

In addition, it still remains to be a challenge to predict the high-performing features based on human judgement. For example, CSIFT base classifier performances on E008 ('flash mob gathering') are somewhat counter-intuitive. First,

Table 4 Objects with highest positive weights as learned by Object Bank linear SVM classifier

Event	Objects with highest weights				
E006	Shield	Car	Loudspeaker	Duck	Pen
E007	Aqualung	Keyboard	Hook	Switch	Blind
E008	Monkey	Boot	Chair	Microwave	Floor
E009	Fruit	Aqualung	Baseball	Cabinet	Blind
E010	Motorcycle	Truck	Jersey	Radio	Bathtub
E011	Bench	Pool table	Fork	Bottle	Desk
E012	Snake	Motorcycle	Monitor	Aqualung	Cesspool
E013	Blanket	Snake	Soccer ball	Sail	Flipper
E014	Blanket	Aqualung	Cesspool	Snake	Blind
E015	Bathtub	Shelf	Loudspeaker	Chair	Bookshelf

The list of the highly ranked objects suggest that semantic interpretation of the learned model is challenging

that the level of performance is quite good despite not explicit detecting people. And, second, that the performance of audio features is quite a bit worse than visual features on this event, despite the videos often having prominent musical accompaniment. Again, this observation highlights that it is important to include diverse features.

7.2.1 Analysis of mid-level semantic features

It can be observed that semantic features such as OB or ASMs can provide superior performance over low-level features on certain event classes, e.g., OB for E009 and ASM for E006 as shown in Table 3. Accordingly, the benefits of incorporating semantic features in terms of performance measure is clear.

We analyzed whether the models learned by semantic features provide the desired *interpretability* which can be another crucial benefit over using low-level features. A meaningful analysis is to look into the types of concepts which are ranked highly for each event class. By analyzing the feature weights learnt by the Object Bank linear SVM classifier, we can determine which concepts in the lexicon are the most important for classifying a given event. Table 4 enlists the five OB objects with highest positive weights, for every event. A high response for one of the positively weighted objects provides evidence in favor of classifying a clip to the corresponding event.

We can see that for most of the events, the objects with high positive weights are *not* strongly related to the respective event class. Such observation suggests that while Object Bank base classifiers are valuable in discriminating between events, it is still challenging to directly draw semantic interpretation from the trained models based on semantic concepts. We have conducted a similar analysis on ASM features, and observed a similar pattern.

The most likely reason for this is that the accuracy of each object detector in OB feature is too low to analyze the contents in unconstrained videos. For example, the object detectors in OB system are trained from relatively regularized and high-resolution image datasets such as ImageNet [12]. It has been previously reported [11] that the accuracy of the object detectors trained by state-of-the-art methods are still very limited, and it degrades further when the detectors are applied to datasets substantially different from the training data. It seems that this limitation is again observed in our evaluation, perhaps even more strongly due to the unconstrained nature of videos in MED ’11 dataset.

7.2.2 Latent SVMs and discriminative description of retrieved videos

In this section, the various aspects of the LSVM results are analyzed. For the evaluation results for this paper, LSVM model was configured to originally discover $S = 40$ scene classifiers for each event and $K = 5$ high-level scene clusters are latently selected during learning, which is a setting found to work well across diverse events.

In terms of performance, the OB L1 model trained by the Latent SVM (LSVM) method outperforms the purely global OB (L0) base classifiers (both max and average pooling) for most event classes, as illustrated in Table 3. Given that our LSVM formulation in Eq. 2 is an extension of L0 model, and still a linear SVM model, the overall improvement in accuracy showcases significant benefits in using LSVM formulation for MED tasks.

More importantly, we find that salient temporal segments identified by the L1 model correlates highly with the key visual contents for the target events. In particular, because the L1 model localizes the segments with key visual evidence from test video clips, the selected segments can be presented to users as a *discriminative summary description* of the retrieval results. This description provides some transparency into the classifier, and can help to further users’ confidence in the system as well as provide an avenue for further research in developing the system.

Figure 8 shows sample description results using the L1 model for 6 retrieved video clips by our system, across dif-

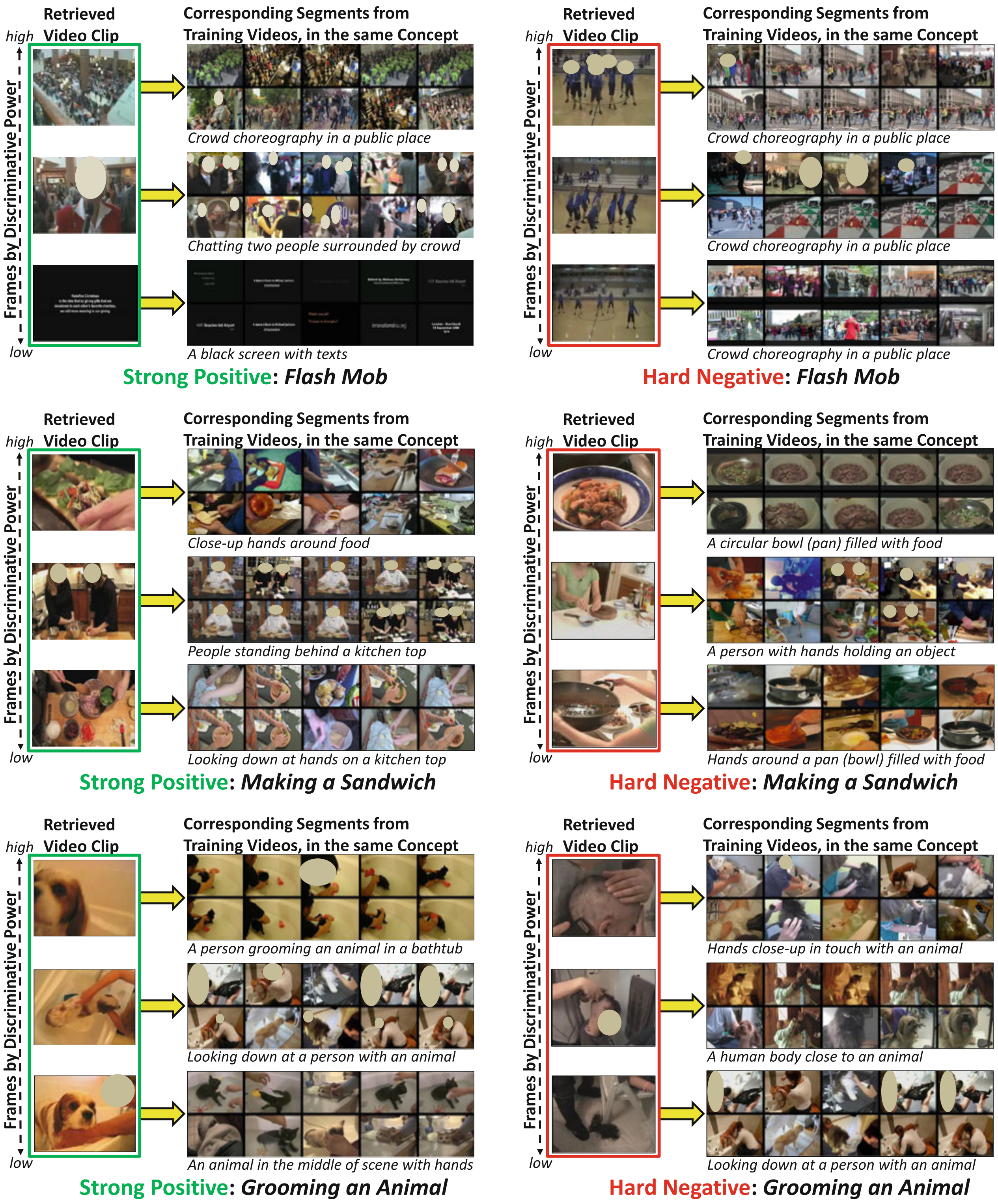

Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
Crowd choreography in a public place
Chatting two people surrounded by crowd
A black screen with texts
Strong Positive: Flash Mob
Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
Crowd choreography in a public place
Crowd choreography in a public place
Crowd choreography in a public place
Hard Negative: Flash Mob
Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
Close-up hands around food
People standing behind a kitchen top
Looking down at hands on a kitchen top
Strong Positive: Making a Sandwich
Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
A circular bowl (pan) filled with food
A person with hands holding an object
Hands around a pan (bowl) filled with food
Hard Negative: Making a Sandwich
Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
A person grooming an animal in a bathtub
Looking down at a person with an animal
An animal in the middle of scene with hands
Strong Positive: Grooming an Animal
Retrieved Video Clip
Corresponding Segments from Training Videos, in the same Concept
high
Frames by Discriminative Power
low
Hands close-up in touch with an animal
A human body close to an animal
Looking down at a person with an animal
Hard Negative: Grooming an Animal

Fig. 8 These results show visualizations of discriminative summary descriptions for retrieved videos, produced by the L1 model. From *left to right*, the three results in the *top half* correspond to three true positive video retrievals for events: the flash mob, making a sandwich, and making a sandwich categories, respectively. The three results in the *bottom half* are hard negatives (i.e., confusers). For each video example, a column of frames is shown on the *left* which shows three (out of five, for brevity) frames that were latently selected by the L1 model as being the most discriminative. On the *right*, the associated scene type clusters corresponding to the latently selected frames are shown. For visualization purposes, the top ten frames that are closest to each concept cluster center are shown, and their visual contents are summarized as text descriptions below each clusters. (Faces are occluded for privacy)

ferent events. The left row shows three *strong positives* (correct retrieval) on three of the test video clips, along with the corresponding targeted events at the bottom. The three examples in the right row correspond to *hard negatives* (incorrect retrieval). For each example, on the left side, there are three frames latently selected and temporally localized by the model as being the most discriminative, which are effectively key evidence or discriminative descriptions for each retrieval. The frames are ordered based on the estimated discriminative power, from top (most discriminative) to bottom. Also shown for each of the localized frames, are the corresponding 10 images from the training data which are in the same concept cluster as the selected latent frame. For better visualization, visual contents for such concept clusters are summarized as text description below each cluster.

The results of discriminative descriptions by the L1 model highlights several benefits of this model. First, it shows that the corresponding segments in the same cluster across training and test data indeed exhibit semantically similar visual content. For example, in the strong positive example for Flash Mob (top-left in Fig. 8), all video segments in the first row include scenes with crowd choreography in a public place, while those in the third row contain black screens with texts. Such consistency showcases that our design of concept detection is effective and generalizes across diverse unconstrained videos, enabling the LSVM framework to successfully operate on a broad class of data. Second, for correct retrieval, the localized subset of frames deliver a concise and effective evidence even when the original video clips are long and cluttered with irrelevant background content. In other words, discriminative frames learnt by LSVM (the left row in each example) seem convincing and related to a target event category. Finally, even for hard negatives, the selected frames do provide a reasonable explanation about its confusion. For example, the hard negative example for *Grooming an animal* in Fig. 8 is actually an event of cutting someone's hair where the selected frames are surprisingly similar to various concept clusters deemed to be discriminative for the target event. This way, the underlying basis for confusion are more transparently understood when the matching segments for the selected segments are visualized altogether, boosting the user confidence on the system as well as providing the valuable information about the system limitations to the system engineers for further research.

Overall, the demonstrated summary description benefits of the developed LSVM framework is significant and provides an avenue for further research. In particular, such description capability is hardly achieved by most base classifiers based on low-level BoW features, and even the OB L0 base classifiers with semantic features do not necessarily support such functionality as discussed in Sect. 7.2.1. However, while the benefits of temporal LSVM model is indeed significant, the overall MED performance is frequently not as high as the OB base classifiers with non-linear SVMs (see Table 3). We believe that additional performance improvement may be achieved by further developing a kernelized version of (linear) L1 model, incorporating techniques such as our recent work [58], which remains as future work.

7.3 Score fusion

As mentioned in Sect. 6.4, we evaluate several fusion methods for all events, and choose the model which produces the best results for each. Note that the fusion model is chosen using base classifier performance on a validation set constructed from approximately 100 positives for each event class, combined with negatives from a development set (the TRECVID event kits and DEV-T collections, respectively). Table 3 shows the base classifiers' performance on our test set, which is different than the performance used for fusion training and model selection. The results of the fusion process is shown in Table 3 and Fig. 9. Over the 10 events, fusion reduces the probability of missed detection by 8 % compared to the best of the base classifiers.

An interesting question is whether the fairly large number of base classifiers that we use are really necessary. There are two ways to answer this. The first is to check whether the learned fusion models assigning a weight of zero to base classifiers; in practice, we find that this is rare, though many base classifiers are given a small weight. The other way to

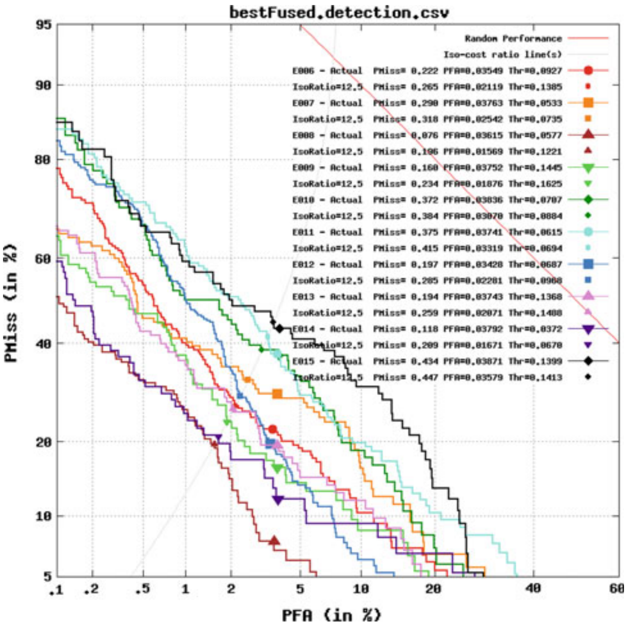


Fig. 9 Fusion results on MED11 test data. These DET curves show the overall system performance after the fusion of 21 base classifiers on the 10 MED11 events. Each *curve* shows performance for different target event where the *x*-axis corresponds to probability of false alarm, and *y*-axis indicates probability of miss, i.e., *lower left corner* corresponds to perfect retrieval

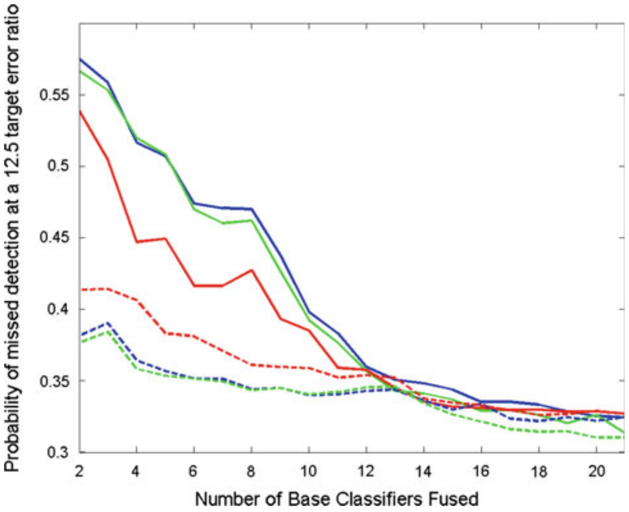


Fig. 10 Comparing the progression of fusion. *Curves* show the average probability of missed detection over the 10 events (lower is better) as a function of the number of base classifiers fused using different methods. The *blue curves* show the result of arithmetic average-based fusion, the *red curves* geometric mean, and the *green curves* show LEF fusion. *Dashed lines* show the progression when the best base classifiers are fused first, and *solid lines* when the worst base classifiers are fused first (color figure online)

evaluate the utility of these base classifiers is to test fusion performance as a function of the number of base classifiers combined. Figure 10 shows the average event performance as a function of the number of base classifiers fused, using three different fusion methods: arithmetic mean (blue curve), geometric mean (red), and LEF (green). For the dashed lines, we start with the best base classifier and, at each step, add the next best; for the solid lines, we start with the worst base classifier and add the next worst. So, at a particular position on the horizontal axis (e.g., 4), the red dashed curve represents the fusion of the 4 best base classifiers, and the solid curve the fusion of the 4 worst. While these curves have their highest gradient magnitude over small numbers of base classifiers, extrapolating the solid lines suggests that there may still be improvement from adding additional high-quality base classifiers.

These results suggest a few further conclusions. First, we could generally achieve the benefits of the learned fusion methods such as LEF over the untrained fusion methods when given enough number of base classifiers, 12 and 13 for fusion from the worst (solid) and the best (dashed) base classifiers, respectively. In contrast, when small number of scores are used for fusion, LEF showed similar, if slightly better, performance to arithmetic mean-based fusion. It implies that the effects of learning a fusion model in the small-dimensional score space may be limited. Second, geometric mean performs much better than the other models when the base classifier set is composed of relatively few poor-performing base classifiers (solid line for 2–10 base classifiers), but significantly worse than the others when the base classifier set is high-performing (dashed lines over the same range). The observed results imply that we can benefit from geometric mean, which tends to be largely affected by lower-valued scores among its inputs, when the high scores from base classifiers are relatively unreliable. On the other hand, average fusion performs well if high scores of base classifiers are reliable.

8 Conclusion

A system for multimedia event detection has been presented in this paper. Various innovations have been integrated into our system, including novel features, Latent SVMs for temporal concept localization, a robust fusion learning scheme, and novel fusion algorithms, among others. The evaluation of the presented system on the large scale TRECVID MED ’11 dataset highlights the benefits of the proposed methods, in terms of detection accuracy along with additional capabilities of discriminative summary descriptions. In particular, our in-depth analysis of the results show that the proposed recounting scheme based on LSVM can effectively help users to understand the rationale for retrievals. On the other hand, our analysis shows that the alternative approach of directly using mid-level concept responses is still challenging.

Acknowledgments This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

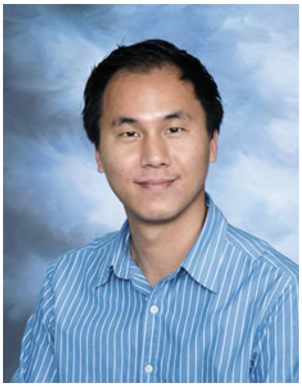
References

1. http://www.lscom.org/
2. TRECVID 2011 Multimedia Event Detection Evaluation Plan Version 3.0. http://www.nist.gov/itl/iad/mig/upload/MED11-EvalPlan-V03-20110801a.pdf
3. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: ICML (2004)
4. Bao, L., Cao, J., Zhang, Y., Li, J., yu Chen, M., Hauptmann, A.G.: Explicit and implicit concept-based video retrieval with bipartite graph propagation model. In: ACM Multimedia (2010)
5. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: ACM SIGIR, pp. 127–134 (2003)
6. Byun, B., Kim, I., Siniscalchi, S.M., Lee, C.H.: Consumer-level multimedia event detection through unsupervised audio signal modeling. In: InterSpeech (2012)
7. Cao, L., Chang, S.F., Codella, N., Cotton, C., Ellis, D., Gong, L., Hill, M., Hua, G., Kender, J., Merler, M., Mu, Y., Smith, J.R., Yu, F.X.: IBM research and Columbia University TRECVID-2012 multimedia event detection (MED), multimedia event recounting (MER), and semantic indexing (SIN) systems (2012)

8. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: ICCV (2007)
9. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
11. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: ECCV (2010)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
13. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
14. Feng, J., Zheng, Y., Yan, S.: Towards a universal detector by mining concepts with small semantic gaps. In: ACM Multimedia (2010)
15. Feng, Y., Lapata, M.: Topic models for image annotation and text illustration. In: NAACL HLT (2010)
16. Gao, S., Wu, W., Lee, C.H., Chua, T.S.: A mfom learning approach to robust multiclass multi-label text categorization. In: ICML (2004)
17. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: TagProp: discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
18. Hauptmann, A.G., Christel, M.G., Yan, R.: Video retrieval based on semantic concepts. Proc. IEEE **96**(4), 602–622 (2008)
19. Hu, W., Xie, N., Li, L., Zeng, X., Maybank, S.J.: A survey on visual content-based video indexing and retrieval. IEEE Trans. Syst. Man Cybern. Part C **41**(6), 797–819 (2011). URL: http://dx.doi.org/10.1109/TSMCC.2011.2109710
20. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **38**(12), 2270–2285 (2005)
21. Jiang, L., Hauptmann, A.G., Xiang, G.: Leveraging high-level and low-level features for multimedia event detection. In: ACM-MM (2012)
22. Jiang, W., Loui, A.C.: Audio-visual grouplet: temporal audio-visual interactions for general video concept classification. In: ACM Multimedia (2011)
23. Jiang, Y.G., Zeng, X., Ye, G., Bhattacharya, S., Ellis, D., Shah, M., Chang, S.F.: Combining multiple modalities, contextual concepts, and temporal matching. In: NIST TRECVID Workshop (2010)
24. Katagiri, S., Juang, B.H., Lee, C.H.: Pattern recognition using a family of design algorithm based upon the generalized probabilistic descent method. Proc. IEEE **86**, 2345–2373 (1998)
25. Kim, I., Lee, C.H.: Optimization of average precision with maximal figure-of-merit learning. In: MLSP (2011)
26. Kim, I., Oh, S., Byun, B., Perera, A.G.A., Lee, C.H.: Explicit performance metric optimization for fusion-based video retrieval. In: ECCV Workshops, no. 3 (2012)
27. Kim, I., Oh, S., Byun, B., Perera, A.G.A., Lee, C.H.: Explicit performance metric optimization for fusion-based video retrieval. In: ECCV Workshop (2012)
28. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. PAMI **20**, 226–239 (1998)
29. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008)
30. Lan, Z.Z., Bao, L., Yu, S.I., Liu, W., Hauptmann, A.G.: Double fusion for multimedia event detection. In: ICME (2012)
31. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical spatio-temporal features for action recognition with independent subspace analysis. In: CVPR (2011)
32. Lee, C.H., Soong, F.K., Juang, B.H.: A segment model based approach to speech recognition. In: ICASSP (1988)
33. Lee, K., Ellis, D.P.W.: Audio-based semantic concept classification for consumer video. IEEE Trans. Audio Speech Lang. Process. **18**(6), 1406–1416 (2010)
34. Li, L.J., Su, H., Xing, E.P., Li, F.F.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)
35. Liu, J., McCloskey, S., Liu, Y.: Local expert forest of score fusion for video event classification. In: ECCV (2012)
36. Ma, A.J., Yuen, P.C.: Linear dependency modeling for feature fusion. In: ICCV, pp. 2041–2048 (2011)
37. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR (2008)
38. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: ECCV (2008)
39. Natarajan, P., Wu, S., Vitaladevuni, S.N.P., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., Natarajan, P.: Multimodal feature fusion for robust event detection in web videos. In: CVPR (2012)
40. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: ICML (2005)
41. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
42. Over, P., Awad, G., Michel, M., Fiscus, J., Antonishek, B., Smeaton, A.F., Kraaij, W., Quéenot, G.: TRECVID 2011—an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2011. NIST, USA (2011)
43. Over, P., Fiscus, J., Sanders, G., Shaw, B., Awad, G., Michel, M., Smeaton, A., Kraaij, W., Quéenot, G.: TRECVID 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2012. NIST, USA (2012)
44. Putthividhya, D., Attias, H.T., Nagarajan, S.S.: Topic regression multi-model latent dirichlet allocation for image annotation. In: CVPR (2010)
45. Reed, J., Lee, C.H.: On the importance of modeling temporal information in music tag annotation. In: ICASSP (2009)
46. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. PAMI **32**(9), 1582–1596 (2010)
47. Scheirer, W., Rocha, A., Micheals, R., Boult, T.: Robust fusion: extreme value theory for recognition score normalization. In: ECCV, pp. 481–495 (2010)
48. Smith, J., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: ICME (2003)
49. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of ACM Multimedia (2006)
50. Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., Sawhney, H.S.: Evaluation of low-level features and their combinations for complex event detection in open source videos. In: CVPR (2012)
51. Terrades, O.R., Valveny, E., Tabbone, S.: Optimal classifier fusion in a non-bayesian probabilistic framework. PAMI **31**(9), 1630–1644 (2009)
52. Tsao, Y., Sun, H., Li, H., Lee, C.H.: An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition. In: ICASSP (2010)
53. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: ICCV (2009)
54. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps (2011)
55. Wang, C., Blei, D.M., Fei-Fei, L.: Simultaneous image classification and annotation. In: CVPR (2009)
56. Wang, Y., Mori, G.: Max-margin hidden conditional random fields for human action recognition. In: CVPR (2009)

57. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: CVPR (2010)
58. Yang, W., Wang, Y., Vahdat, A., Mori, G.: Kernel latent svm for visual recognition. In: Advances in Neural Information Processing Systems (NIPS) (2012)
59. Ye, G., Liu, D., Jhuo, I.H., Chang, S.F.: Robust late fusion with rank minimization. In: CVPR (2012)
60. Zhang, D., Chen, X., Lee, W.S.: Text classification with kernels on the multinomial manifold. In: SIGIR (2005)

Author Biographies



Dr. Sangmin Oh received his Ph.D. and M.S. in computer science from Georgia Institute of Technology and his B.S. in CS from Seoul National University. Dr. Oh's research interests are computer vision, robotic perception, machine learning, and pattern recognition in general. He is particularly interested in developing both theoretical and practical approaches for challenging problems where data is noisy, partially missing, very large, and highly uncertain. He worked on problem domains including video/image understanding, mobile robotics, augmented reality, and wearable computing. Dr. Oh publishes his work in major academic journals and conferences in these areas, and also serves as program committees/reviewers on these venues. Previously, he developed various extensions of probabilistic temporal models, which have been successfully applied to tracking and activity understanding problems for both human and honey bee behavior understanding, for which he was awarded a Samsung research fellowship (2003–2007). Additionally, he was a primary contributor for the GPS-based localization algorithm for mobile platforms (2004); he codeveloped the first affordance learning algorithms for outdoor robot navigation (2005–2007), and developed a fast facial landmark detection method with his collaborators at Microsoft Research (U.S. patent, 2006). Additionally, he has been involved in diverse research programs which involve robotic visual localization and mapping, dynamic augmented earth visualization project (featured in mainstream media such as CNN, Wired etc, 2008), and vision-based child healthcare (2009). Recently (2010–present), his work has been focused on large-scale multimedia retrieval and complex event recognition from large video collections. He has served as chief scientist for various programs supported by DARPA/IARPA/ONR and enjoys collaborating with academic institutions.

Scott McCloskey is a principal research scientist at the Honeywell ACS Labs. He received his B.S. from the University of Wisconsin-Madison in 1999, M.S. from the Rochester Institute of Technology in 2002, and Ph.D. from McGill University in 2008. Dr. McCloskey's research interests include computer vision and computational photography, as applied to real world recognition problems. He has served on the program committee and as a reviewer for the major conferences and journals in the area.

Ilseo Kim is a research staff at the computer vision team of Kitware Inc., NY. He received his Ph.D. and M.S. degrees in electrical and computer engineering from Georgia Institute of Technology in 2008 and 2013, respectively. He also received his B.S. degree in electrical engineering from Seoul National University in 2006. His research interests broadly include pattern recognition in multimedia, computer vision, machine learning, and image and video signal processing.

Arash Vahdat is currently a Ph.D. candidate in the School of Computing Science at Simon Fraser University, Canada. He received his M.Sc. degree in computer science from the same school in 2011 and the B.Sc. degree in information technology from the Computer Engineering Department, Sharif University of Technology, Iran, in 2009. His research interests are in computer vision and machine learning with a focus on event recognition in unconstrained internet videos.

Kevin J. Cannons received his B.Sc. (honours with distinction) degree in computer engineering from the University of Manitoba, Canada, in 2003, the M.ASc. degree in computer engineering from the University of Toronto, Canada, in 2005, and the Ph.D. degree in computer science from York University, Canada, 2011. Currently, he is a postdoctoral researcher in the School of Computing Science at Simon Fraser University. His major field of interest is computer vision with specific emphasis on visual tracking and spatiotemporal analysis.

Hossein Hajimirsadeghi is currently a Ph.D. student at the School of Computing Science, Simon Fraser University, Canada. He received his B.Sc. and M.Sc. in electrical engineering (Control) from University of Tehran, Iran in 2008 and 2010, respectively. His research interests are in machine learning and computer vision, including multi-instance learning and video content analysis.

Greg Mori received the Ph.D. degree in computer science from the University of California, Berkeley in 2004. He received an Hon. B.Sc. in computer science and mathematics with high distinction from the University of Toronto in 1999. He is currently an associate professor in the School of Computing Science at Simon Fraser University. Dr. Mori's research interests are in computer vision, and include object recognition, and human activity recognition, and human body pose estimation.

Dr. A. G. Amitha Perera is an assistant director of computer vision at Kitware, with research interests in motion segmentation, single and multiple object tracking, long duration tracking, motion pattern learning, activity modeling and recognition, and content-based video retrieval. He is the PI of a number of programs at Kitware on multimedia retrieval, tracking, activity recognition, and video compression. He regularly serves on program committees for computer vision conferences (e.g. CVPR, ICCV, ECCV, ICPR, AVSS, WACV, AVSS).

Megha Pandey works as an R&D engineer in the computer vision group at Kitware Inc. She received an M.S. in computer science from the University of North Carolina at Chapel Hill in 2011. Her research interest include computer vision and pattern recognition in images and videos.

Jason J. Corso is an assistant professor in the Computer Science and Engineering Department of SUNY at Buffalo. He received his Ph.D. in computer science at the Johns Hopkins University in 2005. From 2005–2007, Corso was a postdoctoral research fellow in neuroimaging and statistics at the University of California, Los Angeles. He is the recipient of the Army Research Office Young Investigator Award 2010, NSF CAREER award 2009, SUNY Buffalo Young Investigator Award 2011, a member of the 2009 DARPA Computer Science Study Group, and a recipient of the Link Foundation Fellowship in Advanced Simulation and Training 2004. He holds the Associate Editor position of Computer Methods and Programs in Biomedicine since 2009. Corso has authored more than 80 papers on topics of his research interest including computer vision, medical imaging, robotics, computational biomedicine, machine intelligence, statistical learning, perceptual interfaces and smart environments. Corso is a member of the IEEE, ACM, and AAAI.