Semantic Model Vectors for Complex Video Event Recognition

Michele Merler, *Student Member, IEEE*, Bert Huang, Lexing Xie, *Senior Member, IEEE*, Gang Hua, *Senior Member, IEEE*, and Apostol Natsev

***Abstract*—We propose semantic model vectors, an intermediate level semantic representation, as a basis for modeling and detecting complex events in unconstrained real-world videos, such as those from YouTube. The semantic model vectors are extracted using a set of discriminative semantic classifiers, each being an ensemble of SVM models trained from thousands of labeled web images, for a total of 280 generic concepts. Our study reveals that the proposed semantic model vectors representation outperforms—and is complementary to—other low-level visual descriptors for video event modeling. We hence present an end-to-end video event detection system, which combines semantic model vectors with other static or dynamic visual descriptors, extracted at the frame, segment, or full clip level. We perform a comprehensive empirical study on the 2010 TRECVID Multimedia Event Detection task (http://www.nist.gov/itl/iad/mig/med10.cfm), which validates the semantic model vectors representation not only as the best individual descriptor, outperforming state-of-the-art global and local static features as well as spatio-temporal HOG and HOF descriptors, but also as the most compact. We also study early and late feature fusion across the various approaches, leading to a 15% performance boost and an overall system performance of 0.46 mean average precision. In order to promote further research in this direction, we made our semantic model vectors for the TRECVID MED 2010 set publicly available for the community to use (http://www1.cs.columbia.edu/~mmerler/SMV.html).**

***Index Terms*—Complex video events, event recognition, high-level descriptor.**

I. Introduction

RECENT statistics show that videos "in the wild" are growing at a staggering rate [1], [2]. This has posed great challenges for data management, and has attracted the interest of the multimedia analysis research community. These videos are often taken by consumers in unconstrained environments with different recording devices, including cell phones, cameras, camcorders, or professional equipment. They usually contain significant visual variations, largely due to different settings, contents, unconstrained camera motions, production styles, and compression artifacts, to name a few. A challenging task is to build an intelligent system to automatically recognize and detect interesting video events, which will greatly facilitate end users to better index and search video content.

Nevertheless, such videos present new challenges for event detection, not only because of their content diversity and lack of structure, but also for their remarkable growing quantity, which necessitates scalable solutions in terms of both computational cost and memory consumption. For example, on Youtube alone, 35 hours of video are uploaded every minute [2], and over 700 billion videos were watched in 2010[1].

Most previous work in event detection has been on visual surveillance scenarios. Events in videos have been defined in the literature as unusual occurrences in surveillance feeds, such as temporally varying sequence of sub-events [13], [34], motion relativity of visually similar patches [50], or short human actions, which may be modeled by graphical models such as hidden Markov models (HMM) or conditional random fields (CRF) [14], [53], [61], [63]. Usually, the event only exists for a short time span of up to a few seconds. Only recently, people have started working on complex video event detection from videos taken from unconstrained environment, where the video events consist of a long sequence of actions and interactions that last tens of seconds to several minutes. A number of previous works model such complex events as combinations of actions [37], scenes [35], people, and objects [36].

Related research has explored various visual features, either static or spatiotemporal, to serve as the visual representation for recognition, as documented in the TRECVID benchmark [43] of past years. For example, local features [22], [54] extracted at spatiotemporal interest points [8], [23], [54] have been shown to obtain the best recognition accuracy for the task of action recognition [51]. Most of these approaches attempted to model the visual concepts or events directly from the low-level features [5], [25], [62], [65]. Notwithstanding their demonstrated success, we believe that for complex events, an intermediate level semantic representation will help bridge the semantic gap between events and low-level features. Such intermediate semantic representation will make use of discriminative learning to account for the large variations in low-level features that correspond to the same semantic concept (e.g., "people", "cake"), and allow

[1]http://www.youtube.com/t/press_statistics

Manuscript received January 27, 2011; revised July 10, 2011; accepted August 26, 2011. Date of publication September 22, 2011; date of current version January 18, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qingshan Liu.

M. Merler is with the Department of Computer Science, Columbia University, New York, NY 10027 USA (e-mail: mmerler@cs.columbia.edu).

B. Huang is with the Department of Computer Science, University of Maryland, College Park, MD 20740 USA (e-mail: bert@cs.umd.edu).

L. Xie is with the Research School of Computer Science, The Australian National University, Canberra, ACT 0200, Australia (e-mail: lexing.xie@anu.edu.au).

G. Hua is with the Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: ghua@stevens.edu).

A. Natsev is with the IBM TJ Watson Research Center, Hawthorne, NY 10532 USA (e-mail: natsev@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2168948

events to be expressed directly in terms of concepts, rather than low-level features (e.g., "baking a cake" would contain at least one person interacting with cake materials in various stages).

The semantic model vectors we adopted form such an intermediate semantic layer, composed by concatenating the output from 280 discriminative semantic detectors. Based on our prior work [10], [30], [44], each of these semantic detectors is an ensemble SVM trained from thousands of labeled web images. These semantic descriptors cover scenes, objects, people, and various other visual semantics. Each of these semantic dimensions provides the ability to discriminate different semantics from low-level and mid-level visual cues, even though such discrimination may be noisy and imperfect. We previously proposed the semantic model vector representation for semantic content-based multimedia retrieval and mining [30], [44] as well as for modeling the dynamic evolution of semantics within video shots [10]. This paper is a natural extension for clip-level event modeling and detection. Torresani *et al.* [47] followed the same intuition and proposed "classemes" for object category recognition, training weak object classifiers based on an ontology of visual concepts, while Xu *et al.* [57] used the Columbia374-baseline [59] semantic concept classifiers for video event classification. As an alternative way to generate a mid-level semantic representation, unsupervised topic modeling such as LDA/PLSA on top of bag-of-visual-word local features has been employed to model video events [52]. These works differ substantially from the proposed semantic model vectors, which are discriminative classifiers with explicit assigned meanings and learned from a completely separate and independent training set.

We adopt the dataset released by the TRECVID Multimedia Event Detection (MED) Track in 2010 for evaluating our semantic model vector-based representation. This dataset contains 3778 video clips, representing three complex events, i.e., *Assembling a shelter*, *Baking a cake*, and *Batting in a run* plus a *random* category with videos not representing any of the other three categories. Each of these events spans an entire video clip (up to one hour) and is comprised of a collection of different objects, scenes, and actions. For example, the event *Batting in a run* involves a hitter batting, followed by a baserunner scoring a run.

We carefully examined the performance of our proposed semantic model vectors on the TRECVID MED task, and compared it with other low-level visual descriptors, which include both static and spatiotemporal visual features. We built SVM classifiers on top of our semantic model vectors, and on top of static visual (global or local) and dynamic spatio-temporal descriptors, extracted either at frame (keyframes or temporally sampled frames) or video level. Our study revealed that the proposed semantic model vector-based representation produced the best classifier over all detectors based on a single descriptor for this video event detection task, which achieved an average precision of 0.392. This demonstrated that the semantic model vectors indeed can better bridge the semantic gap.

Our investigation also implies that the semantic model vectors are largely complementary to low-level visual features. Therefore, we further develop a comprehensive solution for video event detection by fusing the semantic model vectors with low-level visual features. In our solution, both early and late feature fusion is performed in a hierarchical fashion, which groups together both static and dynamic feature classifiers. Our empirical evaluation indicates that such fusion can significantly boost the recognition accuracy, i.e., it increases average precision to 0.46 for detecting the target video events.

The remainder of the paper is organized as follows: in Section II, we review related work on complex video event detection, Section III describes in detail the type of video events investigated in this work. We present details of the semantic model vectors in Section IV. In Section V, we introduce the proposed framework with the features used for event recognition, also exploring feature fusion strategies. We report the results of our experiments on the 2010 TRECVID MED corpus and discuss them in Section VI. Finally, in Section VII, we draw conclusions and discuss future research directions.

II. Related Work

We will briefly discuss some related work on complex video event detection. We will first summarize some previous technical efforts, followed by a discussion of related benchmark datasets for evaluation.

Largely inspired by the success of structure-free bag-of-words (BoW) representations for visual object recognition, previous systems have employed bags of visual words, or spatiotemporal visual words, to describe objects, scenes, actions, or events [40]. For example, Ballan *et al.* [4] represent an event as a sequence of BoW histograms and use a string kernel to match them. Zhou *et al.* [64] use Gaussian mixture models (GMM) instead of the standard BoW approach to describe an event as a SIFT-Bag. Jiang *et al.* [19] follow the same direction by defining an event as a combination of short-term audio-visual atoms. Nevertheless, valuable spatial context is neglected due to the spatial structure-free BoW representation, which limits the potential of these methods. A more global approach has also been pursued to model directly a whole scene employing holistic, biologically inspired descriptors such as GIST [33] and its derivatives [15], [16], [45].

Besides direct discriminative modeling, some works leverage web data (images [17] or videos [9]) to build reference models for actions or events. A few attempts have also been made to build and leverage ontologies to describe events [3], [6], [42]. Some other works have focused on context and interactions among objects, scenes, and people in order to describe and recognize complex events. Contextual information indeed has been modeled in many different ways, such as human-object interaction [46], [55], [60], visual context for object [11] and scene recognition [56], scene and action combination [29], object and action combination [12], and object, person, and activity relations [38]. Moreover, temporal context has been explored in previous work by modeling complex activities as series of simple actions [21], [39]. Some works model concepts in atomic units (action, scene, and object) for still images [24] or short, simple actions [18].

Before the TRECVID 2010 MED task dataset, there was no clear benchmark for complex video event detection. A few datasets have been introduced in previous work for simple

TABLE I
STATISTICS ON THE CURRENTLY AVAILABLE VIDEO EVENT DATASETS. NOTE HOW THE 2010 TRECVID MED CORPUS CONTAINS MUCH LONGER VIDEOS THAN ANY OTHER AVAILABLE DATASET

Dataset	# Videos	Avg. Length(sec)	# Classes
KTH	559	12.5	6
Hollywood2	2,517	14.8	12
UCF50	6,680	7.44	50
Kodak	3,231	97	25
TRECVID 2005	45,000	12.8	14
CCV	9,317	80	20
TRECVID MED	3,778	**121.27**	3

action recognition, including KTH [41], Moving People Stanford [32], Hollywood [22], Youtube actions dataset [26] (now expanded to UCF50), and UCF Sports Actions dataset, to list a few. INRIA recently introduced the Hollywood2 [29] dataset, which explicitly aims at finding a correlation between actions and scenes. However, the only existing datasets which may be suitable for evaluation of research on complex video events are the Kodak's consumer video benchmark [27] and the TRECVID 2005 news collection. Most recently, the Columbia Consumer Video (CCV) dataset [20] was introduced, which is also designed for event category recognition in unconstrained videos.

Our attention is on videos "in the wild", where a typical video is Youtube-like, often generated by users in unconstrained environments. Some datasets present the same "wilderness" of unconstrained Youtube videos in terms of lack of editing, non-professional recording, and variety of illumination, camera motion, background clutter, changes in object appearance, etc. However, as reported in Table I, most of the existing sets are quite limited in length, particularly when compared to the average length of Youtube videos, which is approximately 4 min and 12 s.[2] As mentioned in Section II, the TRECVID MED corpus we investigate in this work is the closest to Youtube in terms of clip length.

The definitions of the video events we analyze, as posted on the TRECVID MED official site,[3] are the following:

- *Assembling a shelter*: One or more people construct a temporary or semi-permanent shelter for humans that could provide protection from the elements.
- *Batting in a run*: Within a single play during a baseball-type game, a batter hits a ball and one or more runners (possibly including the batter) scores a run.
- *Making a cake*: One or more people make a cake.

Some example video frames from each category are reported in Fig. 1. Clearly evident are the substantial challenges imposed by the widely varying content, quality, viewpoints, settings, illumination conditions, compression, and so on, both across and within event categories.

Most existing approaches and datasets for video event recognition focus on building classifiers for short actions through spatio-temporal features and for concepts based on low-level visual descriptors extracted from keyframes.

[2]http://www.sysomos.com/reports/youtube#video-statistics

[3]http://projects.ldc.upenn.edu/havic/MED10/EventKits.html

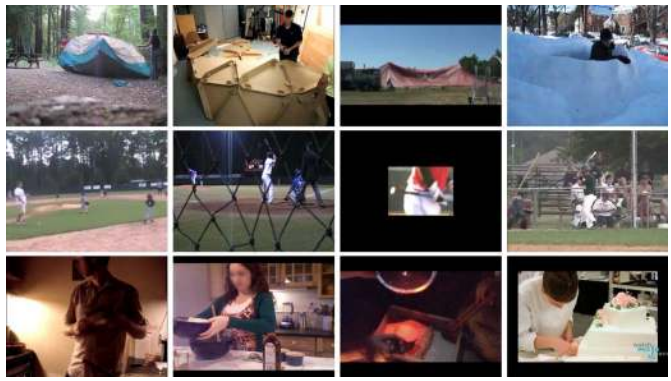

Fig. 1. Examples from the 2010 TRECVID MED video event categories: (top) *assembling shelter*, (middle) *batting in a run*, and (bottom) *making cake*. Note the significant variety of quality, viewpoint, illumination, setting, compression, etc. both intra and inter categories.

III. COMPLEX VIDEO EVENTS

Complex video events however cannot be described by a single action or keyframe. For example, the category *Making a cake* comprises multiple actions (mixing, inserting into oven, tasting) which involve different interactions between primitive semantics (people and objects) over extended periods of time. Therefore, a complex representation which involves such semantic primitives is needed. This concept is illustrated in Fig. 2, where a *Mixing* clip from the UCF50 dataset (top) is compared against a *Making a cake* video from TRECVID MED (bottom), and a subsegment (middle) of the latter. From the significant frames reported for each clip, the greater length and complexity of the bottom *Making a cake* video emerges, since it spans a number of objects and settings (from mixing in the bowl to cooking in the oven, and finally decorating).

A portion of the clip, expanded in the middle row of the figure, represents a part of the cake preparation which involves mixing in a bowl. Notwithstanding a significant difference in visual appearance, the circular movement of the mixer closely resembles that of the spoon in the *Mixing* clip in the top row. Therefore, following the principles adopted in the literature for action and event recognition, one can correctly match the two clips by choosing an appropriate representation, for example the histogram of flow (HOF), which focuses in motion rather than on appearance, assuming that the two sequences are properly aligned temporally.

We build a vocabulary of 100 spatio-temporal HOF words, with K-means clustering, from a dataset of 141 positive and 294 negative clips randomly extracted from the UCF50 dataset, using the widely adopted BoW approach. The graphs in Fig. 2 represent the occurrence frequency of HOF codewords in the three clips. The similarity between the HOF BoW descriptor of the UCF50 *Mixing* clip (top, in green) and the subshot from *Making a cake* video (middle, in blue) is quite clear. However, when considering the whole video from the TRECVID MED set (bottom, in red), we see that the distribution of HOF words is quite different. The reason is that the codewords associated with the mixing part are "masked" by the distribution of the codewords associated with the extremely large number of features appearing in the remainder of the clip. Hence, the complexity and length of the video play against its distinctiveness from the

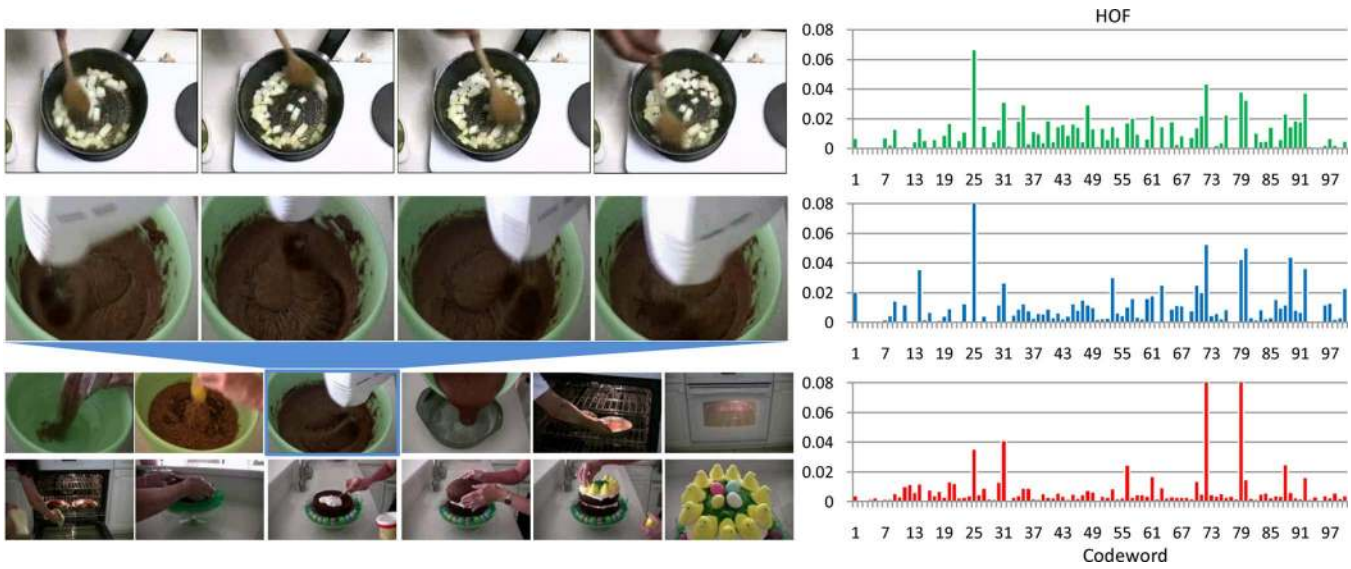

Fig. 2. Comparison between (top, in green) a UFC50 *Mixing* clip and (bottom, in red) a TRECVID MED *Making a cake* video based on 100 codewords bag of words HOF representation. There is a close similarity in codewords distribution between the *Mixing* clip and a subshot of the *Making a cake* clip (expanded from the bottom to the middle row, in blue), which is also evident in the circular motion shown in the frames (on the left side of the figure). Such similarity is lost when the longer, composite whole video in the bottom is considered.

low-level feature point of view. This is also confirmed in the experimental results reported in Section VI-C, where it is clear that the low-level feature representation has drawbacks, even when sophisticated matching schemes are employed, such as temporal pyramid-based matching.

Therefore, the low-level feature-based representations adopted so far in the literature fails to capture the structure and semantics in long, complex video events. In fact, while other types of state-of-the-art low-level descriptors, such as GIST and SIFT, may achieve slightly better performance at the cost of higher dimensionality, they are still limited, as evidenced by the experimental results on the 2010 TRECVID MED collection reported in Figs. 6, 7, and 10.

In this work, we try to alleviate this limitation by using a semantically higher level representation, the semantic model vectors, and integrating it with multiple low-level features in a composite framework. The semantic model vectors, comprising a semantic representation, are able to mitigate the semantic gap, as demonstrated in the experimental results of Section VI.

Besides accuracy, compactness of the descriptors is an important consideration for feasibility, particularly for large-scale multimedia collections such as the one investigated in this work. The proposed semantic model vector, with its 280 dimensions, offers a much more compact representation with respect to traditional low-level descriptors, especially when associated with spatial and/or temporal pyramid frameworks (see the detail in Fig. 7).

In Sections IV and V, we explain in detail the semantic model vectors and the end-to-end event recognition framework.

IV. Semantic Model Vectors

Intuitively, complex temporal events can be described using a combination of elementary visual concepts, their relationships, and temporal evolutions. If an image is worth a thousand words, then a video can be considered the equivalent of a sentence, or even a paragraph.

To this end, we propose an intermediate semantic layer between low-level features and high-level event concepts. This representation, named semantic model vectors, consists of hundreds of discriminative semantic detectors, each derived from an ensemble SVM, trained from a separate collection of thousands of labeled web images, using a common collection of global visual features (described in Section V-B and prior reports [31], [58]). These semantic descriptors cover scenes, objects, people, and various image types. Each of these semantic dimensions provides the ability to discriminate among low-level and mid-level visual cues, even if such discrimination is noisy and imperfect across different data domains.

The semantic model is an ensemble of SVMs with RBF kernel learned from a development set of thousands of manually labeled web images, which were randomly partitioned into three collections: 70% as the training set, 15% as the validation set, and 15% as the held-out set. The number of feature types from which each of the individual SVMs are learned was 98, by means of computing 13 different global visual descriptors including color histogram, color correlogram, color moment, wavelet texture, edge histogram, etc., at up to 8 granularities (i.e., global, center, cross, grid, horizontal parts, horizontal center, vertical parts, and vertical center).

For each feature type, we learn base models (RBF SVMs) from a number N_b of bags of training data, randomly sampled with a balanced number of positive and negative samples, with sample ratio r_d. The default parameters for N_b and r_d used to train all the base models for the semantic models were 2 and 0.2, respectively, which result in a pool of $N = 196$ base models for each concept. To minimize the sensitivity of the parameters for each base model, we choose the SVM parameters based on a grid search strategy. In our experiments, we build the SVM models with different values on the RBF kernel parameters C

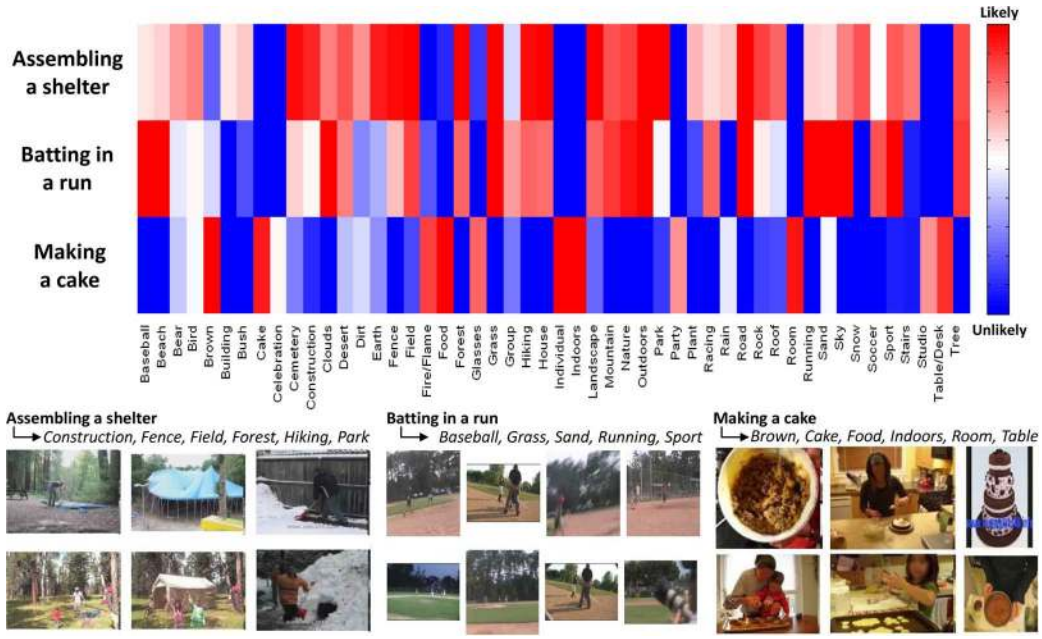

Fig. 3. Usefulness of semantic model vectors for classifying events. (Top) T-score of top 50 model vector dimensions for each event, red indicates positive correlation, blue negative correlation. (Bottom) Example model vectors that are informative for classifying events: $sand \rightarrow batting\_in\_run$ (sand on the beach is similar to sandy baseball courts); $indoors \rightarrow making\_cake$ (cakes are usually cooked in kitchens, whereas assembling shelters and playing baseball are events typically occurring outdoors).

and γ, the relative cost factors of positive versus negative examples, the feature normalization schemes, and the weights between training error and margin. The optimal learning parameters are selected based on the performance measure on 2-fold cross validation on validation data. Each model is then associated with its cross validation performance, where average precision is employed as the performance measure.

Finally, the fusion strategies of the base models into an ensemble classifier are determined based on average precision performance on the held-out data. To reduce the risk of overfitting, we control the strength and correlation of the selected base models by employing a forward model selection step. The algorithm iteratively selects the most effective base model from the unit models pool, adds it to the composite classifier without replacement, and evaluates its average precision on the held-out set. The semantic model output is then the ensemble classifier with the highest average precision observed on the held-out set. This selection step is very fast, and typically prunes more than 70%–80% base models in practice. In fact, the number of selected base models N_i for our 280 semantic models is much smaller than N: the mean and standard deviation values are $N_i = 14 \pm 6$ for $i = 1, \ldots, 280$, where N_i is the number of base models forming the final ensemble SVM for the ith semantic concept (see details in Fig. 4).

Semantic concept detection in a new image x consists in classifying the image using the corresponding ensemble SVM. The score for semantic concept i on x is then

$$SC_i(x) = \sum_{k=1}^{N_i} w_k b_k(x). \quad (1)$$

$SC_i(x)$ is the weighted sum of the individual base models b_k scores on x. The weights w_k are the AP cross-validation scores learned during training.

Semantic concept detection is done starting from the low-level features extracted at multiple, fixed granularities. This process is done at the full image level; no object detection or sliding window approaches are required. The detection time thus depends on the number/type of features selected in the ensemble SVM, plus the extraction time for individual SVM scores. The average extraction time per image is 0.6709 s for the low-level features extraction step, which is shared among semantic models, and 0.0664 ± 0.04 s to obtain the ensemble SVM prediction (mean $\pm$ standard deviation over the 280 semantic models), on a 2.33-GHZ 32-bit Windows machine with 3 GB of RAM.

The average precision and accuracy values for the individual semantic models on the held-out set are, respectively, 0.7579 ± 0.11 and 0.84 ± 0.072 (mean $\pm$ standard deviation over the 280 models). The details of number of base models, extraction time, and average precision scores of each individual semantic model are reported in Fig. 4.

The final semantic model vectors descriptor results from the concatenation of the 280 semantic detectors SC_i for each frame x:

$$SMV(x) = [SC_1(x), \ldots, SC_i(x), \ldots, SC_{280}(x)]. \quad (2)$$

Note that this representation (after being aggregated from frame level to video level) is a lot more compact than most descriptors introduced in Sections V-B and C, as shown in Fig. 7.

Fig. 3 shows a few examples of semantic model vectors. At the top is a visualization of the classification power of model vectors with respect to each of the three target events in the 2010 MED development set.

The x-axis shows a union of the top 50 model vector dimensions most correlated to each event, where correlations scores

Concept	BM	AP	Time	Concept	BM	AP	Time	Concept	BM	AP	Time	Concept	BM	AP	Time	Concept	BM	AP	Time
Airplane	7	0.75	0.033	Door	22	0.9	0.066	Magazine	18	0.86	0.057	SLScreenCapture	16	0.86	0.054	Wire	20	0.75	0.101
Animal	11	0.7	0.058	Downtown	13	0.78	0.169	Magenta	1	0.99	0.005	Ship	7	0.35	0.048	Wood	19	0.86	0.053
Army	15	0.83	0.074	Drawing	15	0.81	0.076	Map	18	0.84	0.006	Shirt	29	0.71	0.145	Yellow	9	1	0.009
BBQ	17	0.89	0.062	Dress	24	0.87	0.092	Market	16	0.87	0.053	Shoe	19	0.74	0.113	Zoo	13	0.88	0.035
BW	1	0.97	0.001	Earth	21	0.71	0.126	Mask	7	0.87	0.028	Sign	8	0.35	0.043	airport	14	0.82	0.056
Baseball_Cricket	13	0.75	0.066	Facecloseup	7	0.72	0.036	Meeting	12	0.4	0.05	Silver	16	0.82	0.085	bear	14	0.7	0.086
Basketball	6	0.31	0.037	Fair	12	0.87	0.039	Menu	13	0.88	0.039	Skiing	7	0.17	0.027	birds	26	0.67	0.139
Beach	19	0.58	0.098	Fashion	13	0.81	0.043	Metal	13	0.78	0.047	Skirt	11	0.59	0.04	boats	15	0.64	0.071
Bed	19	0.85	0.108	Fence	13	0.81	0.042	Microphone	16	0.85	0.05	Skyline_day	10	0.62	0.035	book	20	0.38	0.059
Beer	19	0.79	0.079	Festival	14	0.87	0.066	Mirror	15	0.81	0.062	Skyline_night	24	0.87	0.086	bridge	14	0.38	0.082
Bicycle	16	0.81	0.073	Field	22	0.85	0.093	Monitor	27	0.84	0.121	Smoke	16	0.75	0.06	buildings	19	0.79	0.073
Bird	20	0.73	0.079	Fire_Flame	17	0.86	0.075	Monument	10	0.91	0.035	Snooker	2	0.52	0.01	cars	16	0.48	0.06
Black	1	1	0.002	Fireworks	9	0.94	0.048	Motorcycle	11	0.43	0.03	Snow	13	0.79	0.087	castle	13	0.31	0.06
Blonde	12	0.79	0.041	Fish	14	0.46	0.078	Mountain	1	0.51	0.017	Snow_Scene	11	0.69	0.042	cityscape	22	0.79	0.112
Blue	1	1	0.002	Flag	13	0.83	0.07	Mountain_Scene	13	0.76	0.23	Soccer_football	8	0.87	0.051	clouds	15	0.85	0.094
Boat	20	0.7	0.109	Floor	19	0.85	0.081	Museum	20	0.84	0.234	Sport	17	0.86	0.094	coral	13	0.77	0.04
Brick	18	0.9	0.076	Flower_Scene	15	0.84	0.041	Nature	5	0.98	0.103	Stage	13	0.89	0.06	cow	11	0.46	0.032
Brown	1	0.95	0.007	Food	4	0.98	0.01	Necklace	19	0.75	0.134	Stairs	24	0.86	0.101	dancing	15	0.4	0.067
Brunette	9	0.57	0.058	Forest	18	0.9	0.07	Newspaper	17	0.89	0.141	Star	22	0.82	0.099	earthquake	27	0.51	0.016
Building	4	0.97	0.023	Frost	9	0.71	0.039	Office	11	0.84	0.092	Statue_Sculpture	20	0.83	0.074	elk	11	0.67	0.04
Bush	16	0.76	0.076	Fshing	19	0.82	0.05	Orange	5	0.97	0.039	Store	13	0.87	0.05	fire	8	0.44	0.046
Butterfly	20	0.93	0.05	Fur	18	0.83	0.087	Outdoors	1	1	0.011	Street_Scene	10	0.69	0.065	flags	11	0.47	0.017
CD_DVD	14	0.86	0.068	Glass	21	0.76	0.102	Painting	19	0.83	0.147	Streetart	14	0.9	0.065	flowers	19	0.89	0.079
Cake	13	0.91	0.066	Glasses	12	0.85	0.056	Parade	7	0.01	0.051	Studio	20	0.84	0.073	fox	17	0.49	0.101
Camera	15	0.76	0.063	Gold	14	0.78	0.042	Park	17	0.75	0.116	Suit	17	0.75	0.053	garden	12	0.77	0.078
Capitol	7	0.49	0.021	Graffiti	17	0.86	0.087	Party	4	0.76	0.042	Sunset	7	0.98	0.034	glacier	11	0.52	0.031
Car	9	0.9	0.047	Grass	7	0.99	0.034	Pet	12	0.73	0.084	Surfing	16	0.91	0.051	military	25	0.63	0.127
Carnival	16	0.93	0.071	Green	1	1	0.002	Phone	11	0.86	0.067	Swimming	16	0.73	0.02	moon	9	0.57	0.037
Carpet	20	0.86	0.068	Greenery	1	1	0.003	Photo	23	0.75	0.283	Table_Desk	24	0.89	0.109	nighttime	15	0.83	0.076
Cartoon	14	0.62	0.07	Grey	16	0.8	0.065	Photobooth	17	0.95	0.177	Tan	13	0.71	0.062	ocean	16	0.83	0.07
Cat	10	0.54	0.047	Group_People	13	0.85	0.076	Pillow	12	0.87	0.148	Team	11	0.92	0.041	person	24	0.78	0.071
Celebration	12	0.89	0.06	Harbor	11	0.35	0.036	Pink	13	0.9	0.079	Television	15	0.84	0.051	plane	17	0.73	0.051
Cellphone	18	0.86	0.088	Hat	16	0.83	0.089	Plant	13	0.88	0.153	Tennis	2	1	0.005	plants	17	0.74	0.074
Cemetery	21	0.91	0.056	Hiking	13	0.91	0.04	Portrait	8	0.75	0.116	Tile	20	0.84	0.075	police	17	0.45	0.111
Chair	25	0.85	0.116	Hill	14	0.83	0.05	Poster	23	0.87	0.247	Tour	14	0.77	0.068	protest	19	0.67	0.08
Chart	15	0.45	0.035	Hockey	20	0.95	0.088	Presentation	8	0.7	0.046	Tower	14	0.85	0.035	railroad	19	0.42	0.089
Church	18	0.64	0.06	Horse	13	0.88	0.061	PrintAd	21	0.83	0.144	Town	12	0.75	0.068	reflection	12	0.8	0.095
CivilConstruction	18	0.84	0.055	Horses	10	0.51	0.062	PublicAquarium	13	0.9	0.062	Train	2	0.24	0.006	rocks	13	0.74	0.047
Clock	21	0.78	0.071	Hotel	17	0.74	0.062	Purple	8	0.94	0.033	Tree	10	0.9	0.062	running	22	0.43	0.1
Cloud	3	0.96	0.024	House	10	0.67	0.046	Purse	15	0.65	0.081	Truck	19	0.94	0.094	sky	15	0.84	0.078
Coat	13	0.55	0.053	Icon	19	0.88	0.066	Racing	8	0.94	0.042	Urban_Scene	14	0.87	0.06	soccer	26	0.81	0.065
Color	14	0.77	0.057	Individual	24	0.84	0.205	Radio	16	0.84	0.092	Video	13	0.82	0.06	sports	23	0.57	0.153
Computer	6	0.41	0.027	Indoors	1	1	0.004	Rain	24	0.79	0.113	Village	17	0.84	0.042	statue	20	0.24	0.057
Conference	16	0.9	0.076	Infant	10	0.33	0.05	Rainbow	7	0.34	0.027	Violet	17	0.89	0.041	street	10	0.69	0.053
Couch	13	0.83	0.085	Island	19	0.77	0.057	Rally	15	0.9	0.067	Wall	22	0.8	0.086	sun	10	0.83	0.055
Couple	14	0.81	0.068	Jeans	9	0.76	0.048	Red	5	0.99	0.016	War	9	0.77	0.037	surf	9	0.81	0.012
Crowd	9	0.31	0.011	Jewelry	17	0.83	0.077	Ring	20	0.7	0.095	Watch	19	0.71	0.1	swimmers	18	0.62	0.078
Cruise	12	0.81	0.047	Keyboard	16	0.88	0.046	Road	13	0.77	0.103	Water_Scene	10	0.66	0.051	tattoo	17	0.32	0.084
Curtain	18	0.72	0.077	Knitting	18	0.9	0.094	Rock	12	0.82	0.038	Waterfall	4	0.59	0.013	temple	14	0.36	0.111
Cyan	8	0.98	0.004	Lake	1	0.83	0.001	Roof	18	0.76	0.066	Wave	13	0.88	0.04	tiger	17	0.79	0.057
Cycling	13	0.92	0.041	Lamp	16	0.82	0.042	Room	16	0.88	0.068	Wedding	17	0.82	0.08	toy	16	0.45	0.058
Dance	17	0.9	0.081	Landscape	20	0.87	0.055	Sailing	20	0.87	0.074	White	5	0.98	0.018	valley	21	0.84	0.084
Desert	20	0.84	0.057	Leaf	6	0.55	0.021	Sand	17	0.61	0.12	Whitehouse	13	0.97	0.025	vehicle	15	0.77	0.07
Diamond	21	0.61	0.098	Leather	17	0.73	0.08	Scarf	20	0.81	0.122	Whitewater_Raft	13	0.82	0.043	water	11	0.74	0.061
Dirt	18	0.88	0.075	Lens	20	0.8	0.067	School	15	0.82	0.075	Window	17	0.55	0.062	whales	11	0.71	0.036
Dog	13	0.51	0.095	Logo	13	0.8	0.053	Seat	14	0.76	0.05	Winter_Scene	10	0.67	0.053	zebra	8	0.71	0.013

Fig. 4. Statistics for the 280 model vectors. BM: number of base SVM models used to generate the final ensemble classifiers. AP: average precision cross-validation score obtained during the training process. Time: classifier scoring (i.e., evaluation) time per image (in seconds).

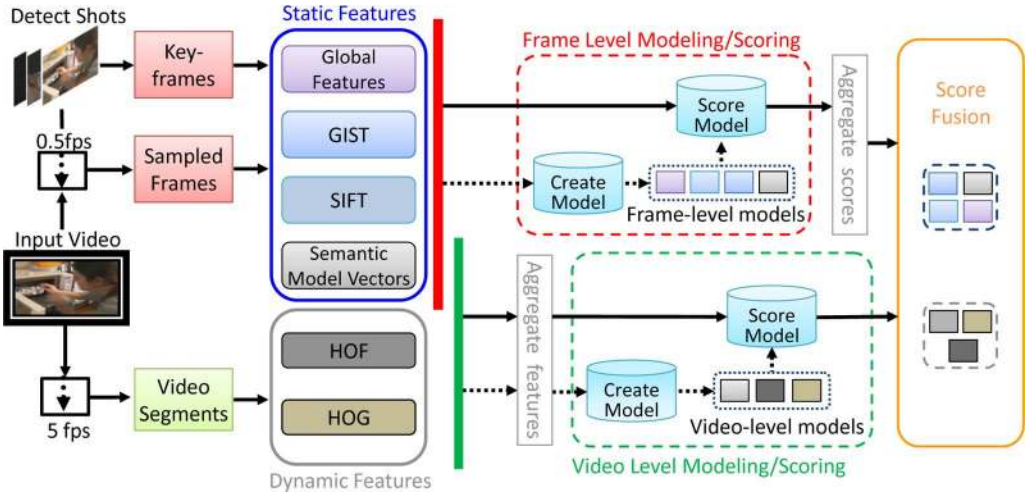

Fig. 5. System framework adopted for video event recognition. We investigated multiple layers of operation/representation: video versus frame level, static versus dynamic features, early versus late aggregation (fusion).

are measured using Student's T-test statistic for score distributions with unequal sample sizes and variances.[4] Red hues denote positive correlation, and blue hues denote negative correlation—likely/unlikely that the concept detector is triggered for the event. The magnitude of the T-scores is reflected in the saturation. Some examples of these concept detector results are included in the bottom part of the figure. Our hypothesis is that the combination and temporal aggregation of the semantic concepts maps closely to complex video events, for example, the *making_cake* event is likely to include *food* prepared on a *table* in an *indoor* setting (i.e., a kitchen). There is not always a one-to-one correspondence between video categories and semantic models, since some of the categories the model vectors were trained on are simply not relevant to the high-level event categories (for example the model vector for Cellphone and the event *Batting in a run*). Nonetheless, if we look at the model vectors outputs as features, the space they span has a dimensionality (280) which is sufficiently high to provide a separation in feature space to discriminate high-level video event categories, even when a direct correlation model vector–high-level event is not found. The SVM models we learn for the high-level video events, using the model vector scores concatenated into a feature vector, basically perform a feature selection.

V. Detection System and Algorithms

Our event detection system includes a variety of approaches. This allows us to explore effective methods for the new multimedia event detection domain, and forms the basis of a comprehensive performance validation.

An overview of the proposed event detection system is shown in Fig. 5, and there are four main parts for processing and learning (roughly from left-to-right in the layout): video processing, feature extraction, model learning, and decision aggregation. The rest of this section will discuss each part in detail.

[4]http://en.wikipedia.org/wiki/Student's_t-test#Unequal_sample_sizes.2C_unequal_variance

A. Video Processing

Each input video is processed a number of different ways in order to extract frame-based and dynamic visual features. Our system has three different modes to prepare a video for feature extraction.

- **Uniformly sampled frames.** We decode the video clip, and uniformly save one frame every two seconds. These frames are later used to extract static visual descriptors: local (SIFT), GIST, Global, and semantic model vectors.
- **Adaptively sampled keyframes.** We perform shot boundary detection using color histogram differences in adjacent frames, and we then take one frame per shot. This frame sampling scheme produces less shots for the event videos since amateur videos tend to have long and unsteady shots. By being temporally adaptive, this scheme may decrease overall appearance diversity in the frames, yet it avoids over-sampling from long shots.
- **Down-sampled short video segments.** We keep short video segments for extracting spatial temporal features (Section V-C). The video sequence is downsampled to five frames per second to reduce computational time, and the spatial temporal features are extracted within windows of four seconds each.

B. Frame Descriptors

We extract a large number of static image features from the sampled frames/keyframes. These features capture a wide range of image information including color, texture, edge, local appearances, and scene characteristics. We build upon these features to extract the semantic model vectors (as described in Section IV) and carry out a comprehensive comparison of state-of-the-art features for classification.

- **Local descriptors** are extracted as SIFT [28] features with dense spatial sampling for keyframes—we use 16 pixels per grid, resulting in approximately 12 000 points per image, and Harris Laplace interest point detection for uniformly sampled frames. Each keypoint is described

with a 128-dimensional vector containing oriented gradients. We obtain a "visual keyword" dictionary of size 1000 (for keyframes) and 4000 (for uniformly sampled frames) by running K-means clustering on a random sample of approximately 300 K interest point features, and we then represent each frame with a histogram of visual words. For keyframes, we used soft assignment following Van Gemert *et al.* [49] using $\sigma = 90$.

- **GIST**: the GIST descriptor [33] describes the dominant spatial structure of a scene in a low-dimensional representation, estimated using spectral and coarsely localized information. We extract a 512-dimensional representation by dividing the image into a 4×4 grid. We also histogram the outputs of steerable filter banks on 8 orientations and 4 scales.
- **Global descriptors** are extracted using 13 different visual descriptors on 8 granularities and spatial divisions. SVMs are trained on each feature and subsequently linearly combined in an ensemble classifier. We include a summary of the main descriptors and granularities. Details on features and ensemble classifier training can be found in our prior report [7].

  Color Histogram—global color distribution represented as a 166-dimensional histogram in HSV color space.

  Color Correlogram—global color and structure represented as a 166-dimensional single-banded auto-correlogram in HSV space using 8 radii depths.

  Color Moments—localized color extracted from a 5×5 grid and represented by the first 3 moments for each grid region in Lab color space as a normalized 225-dimensional vector.

  Wavelet Texture—localized texture extracted from a 3×3 grid and represented by the normalized 108-dimensional vector of the normalized variances in 12 Haar wavelet subbands for each grid region.

  Edge Histogram—global edge histograms with 8 edge direction bins and 8 edge magnitude bins, based on a Sobel filter (64-dimensional).

Having a large diversity of visual descriptors is important for capturing different semantics and dynamics in the scene, as so far no single descriptor can dominate across a large vocabulary of visual concepts and events, and using all has shown robust performance [7], [48].

The spatial granularities include global, center, cross, grid, horizontal parts, horizontal center, vertical parts, and vertical center—each of which is a fixed division of the video frame into spatial layout regions, and then concatenating the descriptor vectors from each region. Such spatial divisions have shown improved performance and robustness in image/video retrieval benchmarks such as TRECVID [43].

C. Spatial-Temporal Features

It is open for debate whether the essence of an event is in its visual appearances over time or in the motion dynamics, though strong arguments for both sides suggest considering both. We compute spatial-temporal features using both motion and dynamic texture information. We detect spatial-temporal interest points (STIP) [23] over the down-sampled video segments (Section V-A), within temporal windows of 20 frames (four seconds). We then compute histogram of gradients (HOG) and histogram of flow (HOF) features extracted in spatio-temporal regions localized around each STIP. For both HOG and HOF features, we generated a codebook of 1000 visual words by clustering a data sample of approximately 300 K and computed bag-of-words histograms similar to those for the SIFT features in Section V-B, with soft assignment.

We explore three aggregation methods both for HOG and HOF. The first is to build a single BoW histogram directly for the entire video, resulting in a 1000-dimensional descriptor [named HOG(F)_Pyr0]. The second employs the temporal pyramid matching scheme [57], with the video temporally split into 2 and 4 segments. A BoW histogram is computed for each shot, and the descriptors are concatenated and weighted according to the temporal level at which they were computed (0.25 for levels 0 and 1, 0.5 for level 2). As reported in Fig. 9, we test two different pyramidal configurations: $\mathrm{HOG(F)\_Pyr1} \times 2$ (3000 dimensional, with whole video and two halves segments concatenated) and $\mathrm{HOG(F)\_Pyr1} \times 2 \times 2$ (7000 dimensional, with whole video, two halves, and four quarters segments concatenated). Since multiple STIP can be detected in the same frame, we also explore computing a BoW histogram for each frame where STIPs were found. We then aggregate from frame level to video level using the same methods employed for the static features and introduced in Section VI, thus obtaining 1000-dimensional vectors. We name descriptors obtained with this third aggregation method simply HOG and HOF.

D. Model Learning

One-versus-all SVMs with RBF kernel are trained, independently for each category, based on each descriptor. During training for one category, all the videos from the other categories (including the random one) are used as negative examples. Parameters C and γ are computed though grid search on a 5-fold cross validation, with a 70% training and 30% validation random splits on both positive and negative examples of the development set. Once the best parameters are determined, the SVMs are retrained on the whole development set.

Either sampling approach seen in Section V-A typically produces multiple frames per video; this yields several features vectors per video for each descriptor (excluding the Pyramid versions of the HOG and HOF features). Given that the problem we investigate consists in classifying whole videos and not individual frames, an aggregation process from frame level to video level is necessary.

We perform such aggregation both at feature level (early fusion) and at prediction level (late fusion). For all features besides Global, the descriptors extracted from the individual frames are combined through average or max aggregation into a single descriptor, representative of the whole video.

We also test aggregation at prediction level, meaning training a classifier at the frame level and then combining the predictions on the individual frames of a test video into a final score. This approach is used for the Global descriptor, for which we take the predictions of the ensemble classifier on the frames of a video and averaged them to obtain the score for the entire video.

Finally, we performed late fusion to combine the predictions of models trained on different descriptors, which offer complementary information. First we grouped static features and dynamic features separately, using linear combinations with uniform weights. We then performed late fusion involving all the descriptors in two ways: hierarchical, as a combination of the static and dynamic sets, and horizontal, as a linear combination of all the features.

E. Feature Fusion

Our baseline approach consists of training RBF kernel SVMs based on individual descriptors. However, we notice that such descriptors are inherently complementary under different perspectives:

- Semantic model vectors operate on a higher semantic level with respect to all the other ones.
- GIST, Global, SIFT, and semantic model vectors are inherently static, as they operate on individual (key)frames, while HOG and HOF are dynamic, as they analyze spatio-temporal volumes within the videos.
- GIST, Global, and semantic model vectors are global features that analyze a whole image, while SIFT, HOG, and HOF model patches localized around local interest points.

Furthermore, even for individual descriptors, we also needed a way of obtaining entire video clip predictions starting from features extracted at multiple frames.

Therefore we apply ensemble late fusion methods to combine all event detection hypotheses generated by the different approaches. We pursue different combination strategies according to the principles adopted:

- Frame-to-video aggregation: for individual descriptors, we try combinations both at the feature level and at the prediction score level, using averaging or max pooling. We find that aggregating frame features into a single video descriptor and obtaining the clip prediction based on the single video-level descriptor outperforms frame-level detection followed by video-level prediction score aggregation.
- Horizontal combination of all descriptors: we average the predictions scores of all individual descriptors at every (key)frame, then perform the same frame-to-video score aggregation described above.
- Hierarchical grouping into static and dynamic features: we first group the descriptors into static (GIST, Global, SIFT, and semantic model vectors) and dynamic (HOG and HOF) and obtain two aggregate frame scores from them (static score and dynamic score). Then, we further combine the static and dynamic scores into a final prediction for every frame and, finally, use average pooling to aggregate from frames to video.

The results of such fusion strategies are reported in Sections VI-B–D.

VI. Experimental Results and Discussion

We evaluate the different approaches outlined in the previous sections on the 2010 TRECVID MED corpus, which consists in 1723 development and 1742 test videos of length varying from a few seconds to an hour, corresponding to three event categories: *Making shelter*, *Baking cake*, and *Batting in a run*, plus a *random* category of videos serving as a distractor set. Table II summarizes the composition of the dataset, specifying also the number of (key)frames extracted with uniform and adaptive sampling. From the table clearly emerges the imbalance between positive and negative examples in both development and test sets.

TABLE II
Video and Frames Distribution of the 2010 TRECVID MED Corpus on Which we Performed Our Experiments

Category	# Videos	# Keyframes	# Sampled Frames
Assembling Shelter	48	2,123	6,931
Batting Run	50	347	2,004
Making Cake	48	3,119	6,292
random	1,577	49,247	95,145
Total Development	1,723	54,836	104,019
Evaluation	1,742	51,306	108,918

Hence, in order to measure the performance of each model, a suitable metric is the AP, which is a popular ranking metric and has been used in all the editions of the TRECVID High Level Feature extraction task. For each category, let TP be the total number of relevant samples in the collection (which contains a total N samples). Let TP_d be the number of relevant samples found in the top d ranked samples returned by the system. Let $I_d = 1$ if the dth sample in the ranked list is relevant and 0 otherwise. The AP is then defined as $1/TP \sum_{d=1}^{N} (TP_d/d) I_d$. The mean of average precision scores over all target categories is defined as the mean average precision (MAP).

In the following, we discuss in detail the results emerging from the experiments.

A. Individual Descriptors Performance

First we compare the performance of event classifiers based on individual descriptors. For all the features, a frame to video aggregation step is performed, as explained in detail in Section VI-B. As reported in Fig. 10, performance varies across categories, with AP scores ranging from 0.15 to 0.3 for *Assembling_Shelter* and *Making_cake*, while *Batting_in_run* ends up being easier to recognize, with AP scores ranging from 0.49 to 0.62. Some general conclusions can be drawn from these MAP scores. Most importantly, the proposed semantic model vectors outperform all the other features in terms of mean average precision (0.392), regardless of which sampling method is used.

From the results presented in Fig. 6 emerges that for any static descriptor, feature extraction on frames obtained by uniform sampling provides better MAP rates than adaptive sampling. This is probably due to the complexity of the events and the "wild" nature of the videos, with a low number of repetitive or static shots. Therefore uniform sampling, which generates a significantly larger number of frames (as shown in Table II), provides richer information to the classifiers.

Considering the large-scale nature of the video event recognition problem at hand, the space occupied by the feature representation of each video is crucial. In Fig. 7 are reported the

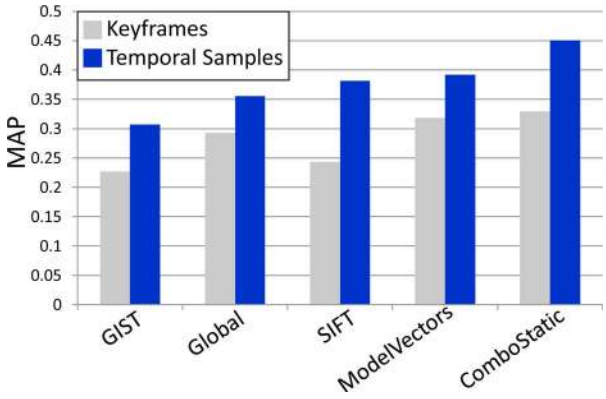

Fig. 6. Mean average precision retrieval performance comparison between keyframes and temporal sampled frames analysis. For each static descriptor, we registered a significant improvement when using temporal sampling, with semantic model vectors resulting as the best descriptor in both cases.

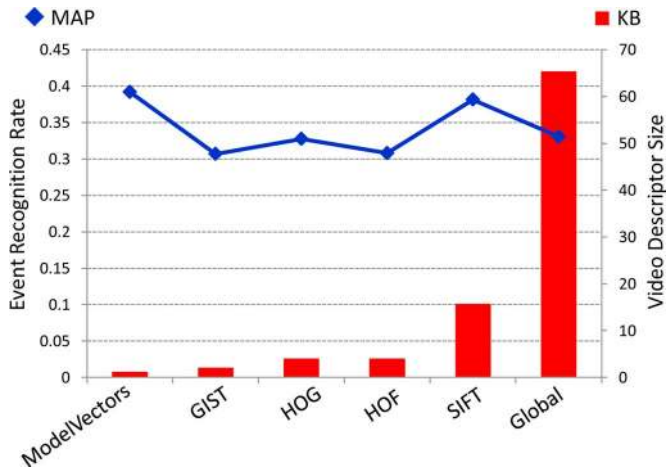

Fig. 7. Mean average precision versus video descriptor size (in kilobytes) based on individual video descriptors. Semantic model vectors offer the most compact representation as well as the best recognition performance.

number of kilobytes (KB) necessary to represent each video (after the feature frames to video aggregation), for each descriptor. Semantic model vectors can represent an entire video with a 280-dimensional feature vector, which is not only the best performing single descriptor but also the most compact one. The second best performing descriptor, the SIFT BoW representation, occupies approximately 15 times more space than the semantic model vector representation. The Global features, comprising multiple descriptors, occupy by far the largest amount of kilobytes.

B. Frame-to-Video Aggregation

As we discussed inSection V-D, since each feature is extracted at frame level, we must aggregate them to determine a single score for each video. We perform an experiment on the semantic model vectors, which is the single best performing descriptor, to determine which aggregation strategy works best. We compare feature-level versus detector score-level aggregation using average or max pooling.

The AP results outlined in Fig. 8 clearly suggest to use feature-level aggregation for all three categories. Hence, we employ this early fusion strategy for all the individual descriptors.

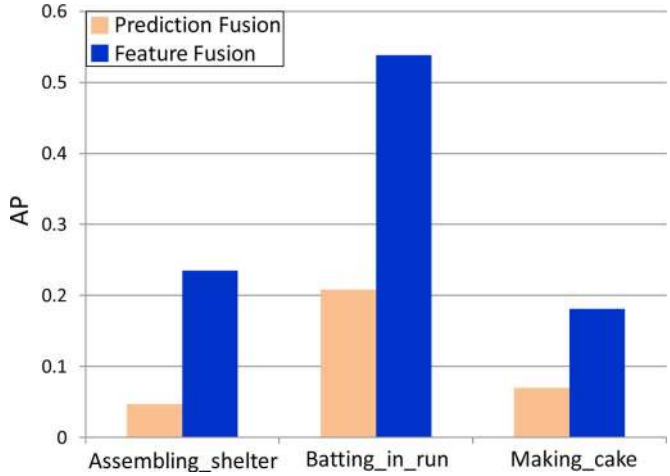

Fig. 8. Semantic model vectors are extracted at every keyframe, thus requiring a fusion from frame level to video level. Fusing features from keyframes into a single descriptor per video and learning a classifier on top of it performs significantly better than learning a classifier directly on the frames and then aggregating the predictions from all the frames into a score for the entire video.

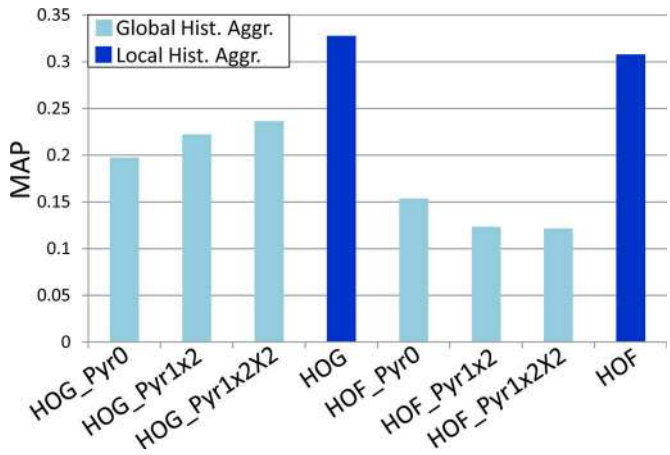

Fig. 9. Mean average precision retrieval performance of the HOG and HOF descriptors: comparison between different methods to generate bags of words video representation. In *global histogram aggregation*, all the visual words from the whole video are binned into a single histogram (plus temporal pyramid declinations), while in *local histogram aggregation*, one histogram per STIP space-time volume is built first, and then all histograms are aggregated using average pooling to form a single video-level descriptor.

The results reported in all other figures in this section follow this framework.

This result corroborates the initial intuition about the complexity of the events we are examining. Classification at the granularity of a single frame, or very a short video segment, is not sufficient to correctly recognize complex video events—a broader context must be inspected instead. Early fusion (or feature-level aggregation) allows each frame to contribute to a richer and more contextualized video representation, therefore providing a more comprehensive description and discrimination for event recognition.

C. Dynamic Features: Global versus Local BoW Histogram Aggregation

As explained in detail in Section V-C, when considering the bag of words approach for spatial-temporal features, there

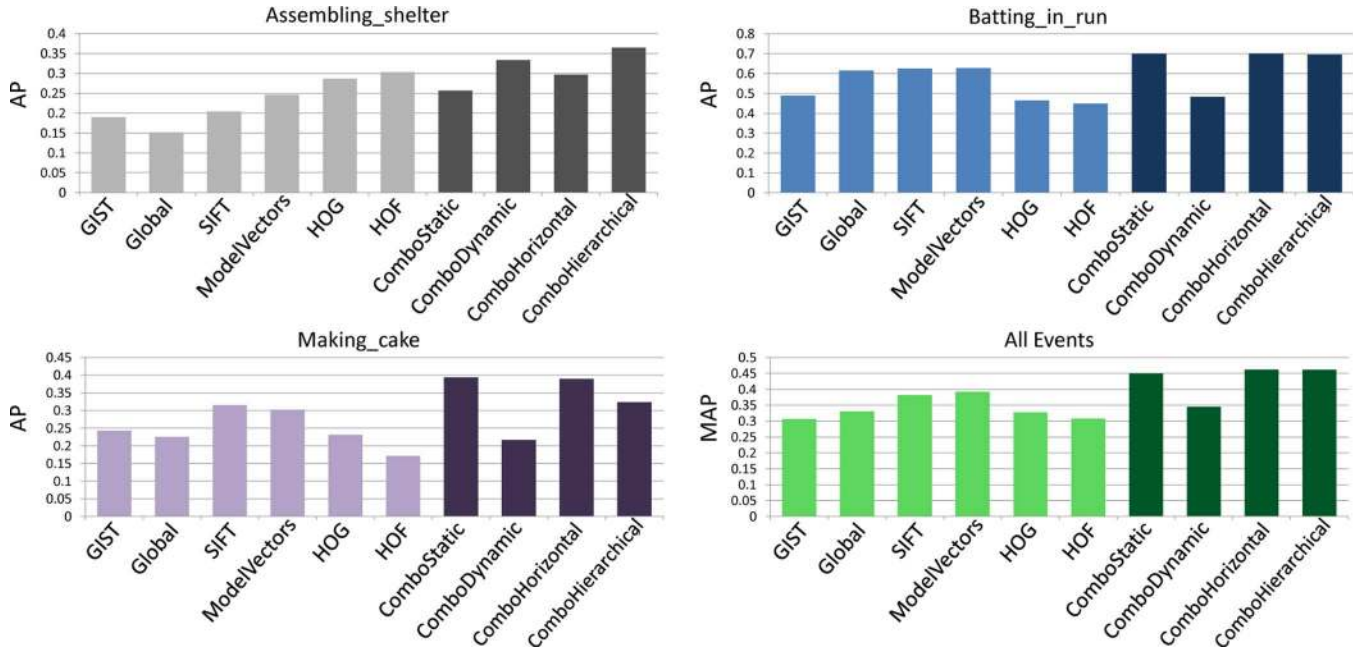

Fig. 10. Retrieval performance of different event recognition approaches based on (lighter colors) individual features and (darker colors) their combinations. Average precision computed for each category and MAP scores over the whole 2010 TRECVID MED dataset.

are different options for building the histogram of codebook words occurrences in a video: one is to bin all interest points descriptors into a single histogram representing the entire video [HOG(F)_Pyr0], which can be further extended to employ a temporal pyramid matching framework to compute separate histograms over temporally partitioned segments ($\mathrm{HOG(F)\_Pyr1} \times 2$ and $\mathrm{HOG(F)\_Pyr1} \times 2 \times 2$). We refer to this strategy as *global histogram aggregation* since all histogram counts are accumulated globally and then normalized. The second option is to generate a separate normalized histogram for each spatio-temporal volume where STIPs have been detected, and then aggregate these local segment histograms into a video-level histogram similarly to the method described in Section VI-B for aggregation of static frame-based features. We call this strategy *local histogram aggregation*, and denote the corresponding runs as HOG and HOF. Note that all dynamic features are video-level features, regardless of the histogram aggregation method.

We compare the MAP performance of the above options in Fig. 9. The results show a significant performance advantage of the *local histogram aggregation* (HOG and HOF bars in dark blue), obtained by averaging BoW local histograms computed from spatio-temporal cubes centered at detected STIP points. This result confirms the intuition expressed in Section III that the target videos are too long and complex to rely on global histograms accumulated over video clips of widely varying lengths. Such representations tend to weaken the contribution of codebook words that may be discriminative locally but whose contribution drowns within the sea of thousands of other (noisy) descriptors accumulated over a long video clip. Local histogram computation and video-level aggregation allows for locally-distinctive codewords to retain a higher weight in the final descriptors, which in turn yields better performance.

In order to alleviate this effect in the pyramid type representations, one might increase their granularity by adding further levels in the temporal pyramid. This idea has two major limitations. The first is the size of the descriptor: already with a pyramid of depth 2, a 7000-dimensional vector is needed (as opposed to the fixed 1000 dimensions of the frame-to-video aggregation used in Fig. 7). The second lies in the hierarchical weight given to higher levels in the pyramid. Finer scale matches are weighted much more than coarse matches. This is desirable if the consecutive, uniformly sampled video sequences describing an event are aligned. The large variety both in appearance and length of the inspected videos suggest that this is not necessarily the case. This could explain why we observe degrading performance for the HOG descriptor as the number of levels in the pyramid increased.

D. Feature Fusion

In Section V-E, we observed the complementary nature of the descriptors from the view point of semantics, temporal, and spatial resolution, which suggests that fusion across these descriptors is likely to improve performance even further.

Here we report the results obtained by late fusion methods to combine all event detection hypotheses generated by the different approaches. We ensured that the scores from all approached were compatible for fusion by applying sigmoid normalization on the non-probabilistic predictors. Fusion was performed by averaging prediction scores. The MAP scores are reported in Fig. 10.

We observe that a combination of static features (ComboStatic, including Global, SIFT, GIST, and semantic model vectors) works better for events that can be predicted by some iconic objects or settings (e.g., a cake for *Making_cake*; the baseball field and player outfits, including helmet and

bat, for *Batting_in_run*), while combining dynamic features (ComboDynamic, including HOG, $\mathrm{HOG\_Pyr1} \times 2$, HOF, and $\mathrm{HOF\_Pyr1} \times 2$ descriptors) performs better for events with more widely varying visual appearance and temporal evolution (e.g., *Assembling_shelter* videos showing different stages of shelter construction and their temporal progression).

Both combinations (or static or dynamic descriptors) boost the average precision scores with respect to the individual descriptors across all three events. The performance behavior of static and dynamic features appears to be complementary across event categories. Hence, we apply a hierarchical fusion (ComboHierarchical), which combines ComboStatic and ComboDynamic predictions. This final fusion step further improves the MAP rate, confirming the complementary nature of static and dynamic features. We also perform an aggregation of all the static and dynamic runs in a single step (ComboHorizontal), and observe performance boost similar to the hierarchical fusion method.

In all the combination cases inspected, late fusion of multiple descriptors results in a boost of MAP with respect to the individual descriptors for all the events in the dataset, thereby confirming the complementary nature of these features. The best MAP performance of 0.46 is achieved by fusing all runs, without an appreciable difference between horizontal versus hierarchical fusion.

VII. Conclusion

We have proposed a system for complex video event recognition in unconstrained real-world consumer videos, such as those from YouTube. Our recognition system incorporates information from a wide range of static and dynamic visual features. We evaluate our framework on the 2010 TRECVID Multimedia Event Detection dataset, a fully annotated unconstrained video collection in terms of content complexity and average video clip length.

In particular, we propose semantic model vectors, an intermediate level visual representation, to help bridge the semantic gap between low-level visual features and complex video events. We use a large number of semantic detectors covering scenes, objects, people, and various image types. The semantic model vector representation turns out to be the best-performing single feature in our experiments, achieving mean average precision of 0.392 on the TRECVID MED10 dataset, while also being the most compact representation (only 280 dimensions versus thousands of dimensions for traditional low-level descriptors). All these properties make this semantic representation particularly suitable for large-scale video modeling, classification, and retrieval. The semantic model vectors also appear to be complementary to the other descriptors. We experimented with different feature granularities (frame-based versus video-based) and fusion types (e.g., early versus late fusion), and observed a performance boost leading to 0.46 overall MAP scores obtained from a late fusion of static and dynamic features.

In the future, we plan to generalize this approach to a wider range of video events. We also plan on learning the temporal evolution of descriptors representing actions, people, objects, and scenes.

Acknowledgment

The authors would like to thank the Associate Editor and the Reviewers for their insightful comments and suggestions, which contributed to improve and make this work complete.

References

[1] "Cisco Visual Networking Index: Forecast and Methodology, 2009–2014," 2010. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/whitepaper_c11-481360_ns827_Networking_Solutions_White Paper.html.

[2] "Great Scott! Over 35 Hours of Video Uploaded Every Minute to Youtube," The official YouTube blog, 2010. [Online]. Available: http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html.

[3] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, "Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies," *Multimedia Tools Appl.*, pp. 313–337, Jun. 2010.

[4] L. Ballan, M. Bertini, A. D. Bimbo, and G. Serra, "Video event classification using string Kernels," *Multimedia Tools Appl.*, vol. 48, no. 1, pp. 69–87, May 2010.

[5] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis. (IJCV)*, vol. 74, pp. 17–31, 2007.

[6] F. Bremond, N. Maillot, M. Thonnat, and V.-T. Vu, "Ontologies for video events," *Research Report Number 5189*, 2004.

[7] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. R. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang, "IBM research TRECVID-2007 video retrieval system," in *Proc. NIST TRECVID Workshop*, 2007.

[8] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2005, pp. 65–72.

[9] L. Duan, D. Xu, W.-H. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] S. Ebadollahi, L. Xie, S. Chang, and J. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, 2006, pp. 881–884.

[11] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Comput. Vis. Image Understand.*, 2010.

[12] A. Gupta and L. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[13] A. Hakeem and M. Shah, "Learning, detection and representation of multi-agent events in videos," *Artif. Intell.*, vol. 171, Jun. 2007.

[14] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2009, pp. 128–135.

[15] Y. Huang, K. Huang, T. Tan, and D. Tao, "A novel visual organization based on topological perception," in *Proc. Asian Conf. Computer Vision*, 2009, pp. 180–189.

[16] Y. Huang, K. Huang, L. Wang, D. Tao, T. Tan, and X. Li, "Enhanced biologically inspired model," in *Proc. Computer Vision and Pattern Recognition*, 2008.

[17] N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff, "Learning actions from the web," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2009.

[18] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2010.

[19] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis, and A. C. Loui, "Short-term audio-visual atoms for generic video concept classification," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2009, pp. 5–14.

[20] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ACM Int. Conf. Multimedia Retrieval (ICMR)*, 2011.

[21] D. Kuettel, M. D. Breitenstein, L. V. Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.
[23] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, no. 2, pp. 107–123, 2005.
[24] L. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Proc. Int. Conf. in Computer Vision (ICCV)*, 2007, pp. 1–8.
[25] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.
[26] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1996–2003.
[27] A. C. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: Concept definition and annotation," in *Proc. ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Sep. 2007.
[28] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[29] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2929–2936.
[30] A. Natsev, M. R. Naphade, and J. R. Smith, "Semantic representation, search and mining of multimedia content," in *Proc. ACM Int. Conf. Knowledge Discovery and Data Mining (KDD'04)*, Seattle, WA, Aug. 2004, pp. 641–646.
[31] A. Natsev, M. Hill, J. R. Smith, L. Xie, Rong, Y. S. Bao, M. Merler, and Y. Zhang, "IBM research TRECVID-2009 video retrieval system," in *Proc. NIST TRECVID Workshop*, 2009.
[32] J. C. Niebles, B. Han, A. Ferencz, and L. Fei-Fei, "Extracting moving people from internet videos," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2008, pp. 527–540.
[33] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[34] K. Prabhakar, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg, "Temporal causality for the analysis of visual events," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.
[35] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 413–420.
[36] P. M. Roth and M. Winter, Survey of Appearance-Based Methods for Object Recognition, 2008, Tech/ Rep. ICG-TR-01/08.
[37] R. W. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, pp. 976–990, Jun. 2010.
[38] M. Ryoo and J. Aggarwal, "Hierarchical recognition of human activities interacting with objects," in *Proc. 2nd Int. Workshop Semantic Learning Applications in Multimedia (SLAM) at CVPR*, 2007, pp. 1–8.
[39] M. Ryoo and J. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2009, pp. 1593–1600.
[40] G. Schindler, L. Zitnick, and M. Brown, "Internet video category recognition," *Internet Vis.*, pp. 1–7, 2008.
[41] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognition (ICPR'04)*, 2004, vol. 3, pp. 32–36.
[42] S. Sen and J. Ma, "Contextualized eventdriven prediction with ontology-based similarity," in *Proc. AAAI Spring Symp. Intelligent Event Processing*, 2009.
[43] A. F. Smeaton, P. Over, and W. Kraaij, "High level feature detection from video in trecvid: A 5-year retrospective of achievements," in *Proc. Multimedia Content Analysis*, 2009, pp. 151–174.
[44] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, 2003, pp. 445–448.
[45] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.
[46] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2004–2011.
[47] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Computer Vision (ECCV)*, Sep. 2010, pp. 776–789.
[48] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[49] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Comparing compact codebooks for visual categorization," *Computer Vis. Image Understand.*, to be published.
[50] F. Wang, Y. Jiang, and C. Ngo, "Video event detection using motion relativity and visual relatedness," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2008, pp. 239–248.
[51] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. British Machine Vision Conf. (BMVC)*, 2009.
[52] J. Wang, L. Duan, L. Xu, H. Lu, and J. S. Jin, "Tv ad video categorization with probabilistic latent concept learning," in *Proc. Multimedia Information Retrieval*, 2007, pp. 217–226.
[53] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and Dulong, "Semantic event detection using conditional random fields," in *Proc. Computer Vision and Pattern Recognition Workshop (CVPR)*, 2006, pp. 109–109.
[54] G. Willems, T. Tuytelaars, and L. J. V. Gool, "An efficient dense and scale-invariant saptio-temporal interest point detector," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2008, pp. 650–663.
[55] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2007, pp. 1–8.
[56] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.
[57] D. Xu and S.-F. Chang, "Video event recognition using Kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, Nov. 2008.
[58] R. Yan, J. Tesic, and J. R. Smith, "Model-shared subspace boosting for multi-label classification," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2007, pp. 834–843.
[59] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, Columbia University's Baseline Detectors for 374 Lscom Semantic Visual Concepts, Columbia Univ., 2007, Tech. Rep.
[60] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.
[61] J. Yin, D. H. Hu, and Q. Yang, "Spatio-temporal event detection using dynamic conditional random fields," in *Proc. 21st Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2009, pp. 1321–1326.
[62] L. Zelnik-Manor and M. Irani, "Statistical analysis of dynamic actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, pp. 1530–1535, 2006.
[63] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted HMMs for unusual event detection," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 611–618.
[64] X. Zhou, X. Zhuang, S. Yan, S. F. Chang, M. H. Johnson, and T. S. Huang, "Sift-bag Kernel for video event analysis," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2008, pp. 229–238.
[65] G. Zhu, M. Yang, K. Yu, W. Xu, and ZYY09, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptors," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2009, pp. 165–174.

Michele Merler (S'09) received the B.S. and M.S. degrees in telecommunications engineering from the University of Trento, Trento, Italy, in 2004 and 2007, respectively, and the M.S. degree from the Department of Computer Science at Columbia University, New York, in 2008. He is currently pursuing the Ph.D. degree in the Department of Computer Science at Columbia University.

He was a visiting student at UCSD for a year in 2005. He has been a Research Intern at the T. J. Watson Research Center for three summers between 2008 and 2010. His research interests are in high-level vision and multimedia processing.

Mr. Merler is a student member of the ACM.

Bert Huang received the B.S. and B.A. degrees from Brandeis University, Waltham, MA, and the M.S., M.Phil., and Ph.D. degrees from Columbia University, New York.

He is a postdoctoral researcher in the Department of Computer Science at University of Maryland, College Park. In the summer of 2010, he was an IBM Research Intern at the T. J. Watson Research Center.

Lexing Xie (SM'11) received the B.S. degree from Tsinghua University, Beijing, China, in 2000, and the M.S. and Ph.D. degrees from Columbia University, New York, in 2002 and 2005, respectively, all in electrical engineering.

She is Lecturer in the Research School of Computer Science at the Australian National University, Canberra. She was with the IBM T. J. Watson Research Center, Hawthorne, NY, from 2005 to 2010. Her recent research interests are in multimedia mining, machine learning, and social media analysis.

Dr. Xie has won several awards: the best conference paper award in IEEE SOLI 2011, the best student paper awards at JCDL 2007, ICIP 2004, ACM Multimedia 2005, and ACM Multimedia 2002. She also received the 2005 IBM Research Josef Raviv Memorial Postdoc fellowship in computer science and engineering.

Gang Hua (SM'11) received the B.S. degree in automatic control engineering and the M.S. degree in control science and engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering at Northwestern University, Evanston, IL, in 2006.

He was enrolled in the Special Class for the Gifted Young of XJTU in 1994. He is currently an Associate Professor of Computer Science at Stevens Institute of Technology, Hoboken, NJ. Before that, he was a research staff member at the IBM Research T. J. Watson Center, Hawthorne, NY, from 2010–2011, a senior researcher at Nokia Research Center, Hollywood, CA, from 2009 to 2010, and a scientist at Microsoft Live Labs Research, Redmond, WA, from 2006 to 2009. He is the author of more than 50 peer reviewed publications in prestigious international journals and conferences. As of September 2011, he holds 3 U.S. patents and has 17 more patents pending.

Dr. Hua is an associate editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and *IAPR Journal of Machine Vision and Applications*, and a guest editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal on Computer Vision*. He is an area chair of the IEEE International Conference on Computer Vision, 2011, an area chair of ACM Multimedia 2011, and a Workshops and Proceedings Chair of the IEEE Conference on Face and Gesture Recognition 2011. He is a member of the ACM.

Apostol (Paul) Natsev received the M.S. and Ph.D. degrees in computer science from Duke University, Durham, NC, in 1997 and 2001, respectively

He is a Research Staff Member and Manager of the Multimedia Research Group at the IBM T. J. Watson Research Center, Hawthorne, NY. He joined IBM Research in 2001. At IBM, he leads research efforts on multimedia analysis and retrieval, with an agenda to advance the science and practice of systems that enable users to manage and search vast repositories of unstructured multimedia content. He is a founding member and current team lead for IBM's award-winning IMARS project on multimedia analysis and retrieval, with primary contributions in the area of semantic, content-based, and speech-based multimedia indexing and search, as well as video fingerprinting and copy detection. He is an avid believer in scientific progress through benchmarking, and has participated actively in a dozen open evaluation/showcasing campaigns, including the annual TRECVID video retrieval evaluation, the CIVR VideOlympics showcase, and the CIVR Video Copy Detection showcase. He is an author of more than 70 publications and 18 U.S. patents (granted or pending) in the areas of multimedia analysis, indexing and search, multimedia databases, and query optimization.

Dr. Natsev's research has been recognized with several awards, including the 2004 Wall Street Journal Innovation Award (Multimedia category) for IMARS, an IBM Outstanding Technical Accomplishment Award in 2005, a 2005 ACM Multimedia Plenary Paper Award, a 2006 ICME Best Poster Award, and the 2008 CIVR VideOlympics People's Choice Award (for IMARS). He is a Senior Member of the ACM.