Machine Vision and Applications (2014) 25:5–15
DOI 10.1007/s00138-013-0529-6

SPECIAL ISSUE PAPER

E-LAMP: integration of innovative ideas for multimedia event detection

Wei Tong · Yi Yang · Lu Jiang · Shoou-I Yu · ZhenZhong Lan · Zhigang Ma · Waito Sze · Ehsan Younessian · Alexander G. Hauptmann

Received: 8 February 2013 / Revised: 3 June 2013 / Accepted: 10 June 2013 / Published online: 9 July 2013

Abstract Detecting multimedia events in web videos is an emerging hot research area in the fields of multimedia and computer vision. In this paper, we introduce the core methods and technologies of the framework we developed recently for our Event Labeling through Analytic Media Processing (E-LAMP) system to deal with different aspects of the overall problem of event detection. More specifically, we have developed efficient methods for feature extraction so that we are able to handle large collections of video data with thousands of hours of videos. Second, we represent the extracted raw features in a spatial bag-of-words model with more effective tilings such that the spatial layout information of different features and different events can be better captured, thus the overall detection performance can be improved. Third, different from widely used early and late fusion schemes, a novel algorithm is developed to learn a more robust and discriminative intermediate feature representation from multiple features so that better event models can be built upon it. Finally, to tackle the additional challenge of event detection with only very few positive exemplars, we have developed a novel algorithm which is able to effectively adapt the knowledge learnt from auxiliary sources to assist the event detection. Both our empirical results and the official evaluation results on TRECVID MED'11 and MED'12 demonstrate the excellent performance of the integration of these ideas.

Keywords Multimedia event detection · Multimedia content analysis

1 Introduction

With ever-expanding multimedia collections, multimedia content analysis is becoming a fundamental research issue for many applications such as indexing and retrieval. One of the interesting problem of multimedia content analysis is to automatically detect some predefined events in a video collection. An event is a complex activity occurring at a specific place and time which involves people interacting with other people and/or object(s) [31]. In general, an event consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have temporal and semantic relationships to the overarching activity. Given a collection of test videos and a list of test events, the task of event detection is to indicate whether each of the test events is present anywhere in each of the test videos. Compared with traditional concept analysis [23,36,38], event detection

W. Tong (✉) · Y. Yang · L. Jiang · S.-I. Yu · Z. Lan · W. Sze · E. Younessian · A. G. Hauptmann
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: tongwei@cs.cmu.edu

Y. Yang
e-mail: yiyang@cs.cmu.edu

L. Jiang
e-mail: lujiang@cs.cmu.edu

S.-I. Yu
e-mail: iyu@cs.cmu.edu

Z. Lan
e-mail: lanzhzh@cs.cmu.edu

W. Sze
e-mail: wts@cs.cmu.edu

E. Younessian
e-mail: ehsa0001+@andrew.cmu.edu

A. G. Hauptmann
e-mail: alex@cs.cmu.edu

Z. Ma
Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
e-mail: ma@disi.unitn.it

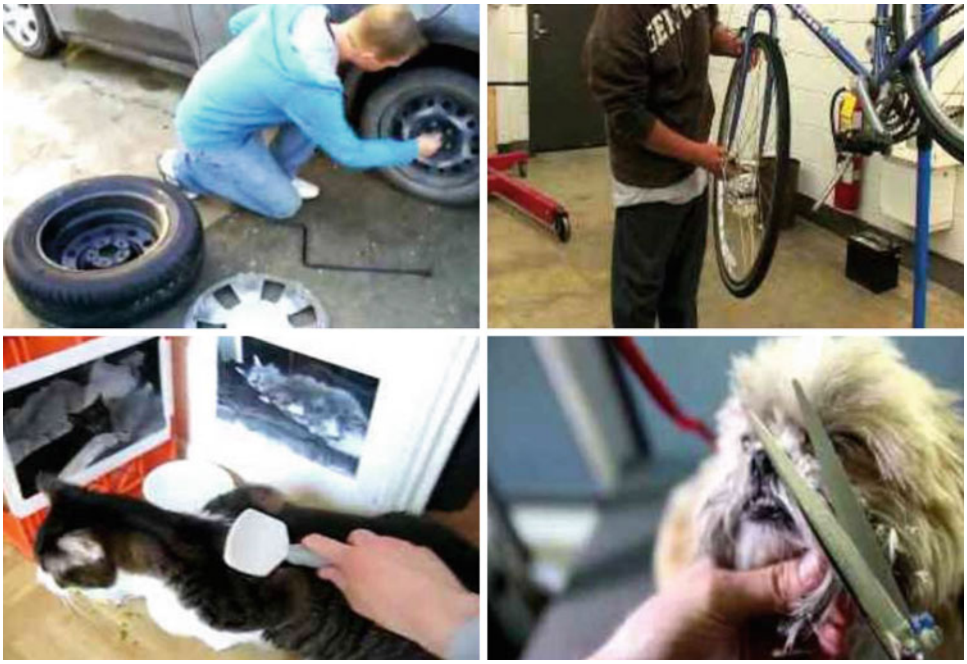


Fig. 1 Example snapshots of the videos from the event "Changing a vehicle tire" and the event " Grooming an animal" defined by TRECVID Multimedia Event Detection 2011 task. The two snapshots in the *first row* are from the event "Changing a vehicle tire" and the two snapshots in the *second row* are from the event " Grooming an animal"

is a more challenging task due to its dynamic attributes and semantic richness. For example, the event of "making a cake" consists of a combination of several concepts such as "cake", "people" and "kitchen" together with the action "baking" within a longer video sequence. Figure 1 shows a couple of example snapshots of the videos from the event "Changing a vehicle tire" and the event " Grooming an animal" which are defined by TRECVID Multimedia Event Detection 2011 task.

The study of the multimedia event detection first emerged in structured scenarios, e.g., surveillance videos, sports videos and news videos [1,33,44,48]. Recently, people started to focus more on general unconstrained videos such as those obtained from internet video sharing web sites like YouTube. To facilitate and encourage the research of new technologies and algorithms for multimedia event detection, the ACM Multimedia society has launched three international workshops on events in multimedia (EiMM'09-'11) and the National Institute of Standards and Technology (NIST) launched the task of "Event detection in Internet multimedia (MED)" in 2010 TREC Video Retrieval Evaluation (TRECVID) workshop [31]. In general, there are three core challenges in detecting multimedia events: The first is that to detect an event one has to extract a comprehensive set of features from the raw video. In general, the procedure of extracting those features is computationally expensive and time-consuming. This is particularly a serious problem for a large collection of videos. For example, in the TRECVID MED'12 task, the testing video set consists of about 4,000 h of videos, and on YouTube there are about 30 million hours of videos uploaded each year. Even with the help of powerful computer clusters, how to efficiently extract a comprehensive set of features over large video collections is still a big challenge. The second challenge is that with extracted features from the videos, what representations should be used so that different aspects of the information conveyed by the features can be effectively utilized to model an event. The third challenge is how to model an event by jointly exploring the multiple modalities provided by either different features or/and different representations of the same features. In addition to these three challenges of general multimedia event detection, a new challenging task was defined by TRECVID MED'12 which is the event detection with few positive exemplars. In this challenge, only a very limited number of positive example videos of an event are provided, specifically,10 positive videos, thus the traditional classification scheme which works well for event detection with relatively large number of positive training examples might not be suitable anymore for the event detection with limited positive examples.

To tackle these challenges of multimedia event detection, we have developed a framework [5,11] within which we implemented the Event Labeling through Analytic Media Processing (E-LAMP) system [5]. Figure 2 shows the overview of the framework. In this paper, instead of describing the whole framework, we focus us on a couple of key components (highlighted by red boxes in Fig. 2) which are novel and essential in helping us to achieve both effective and efficient event detection.

More specifically, for the challenge of efficiency in feature extraction, we conduct comprehensive studies which reveal that using features extracted from the videos with reduced resolution may not degrade the performance of event detec-

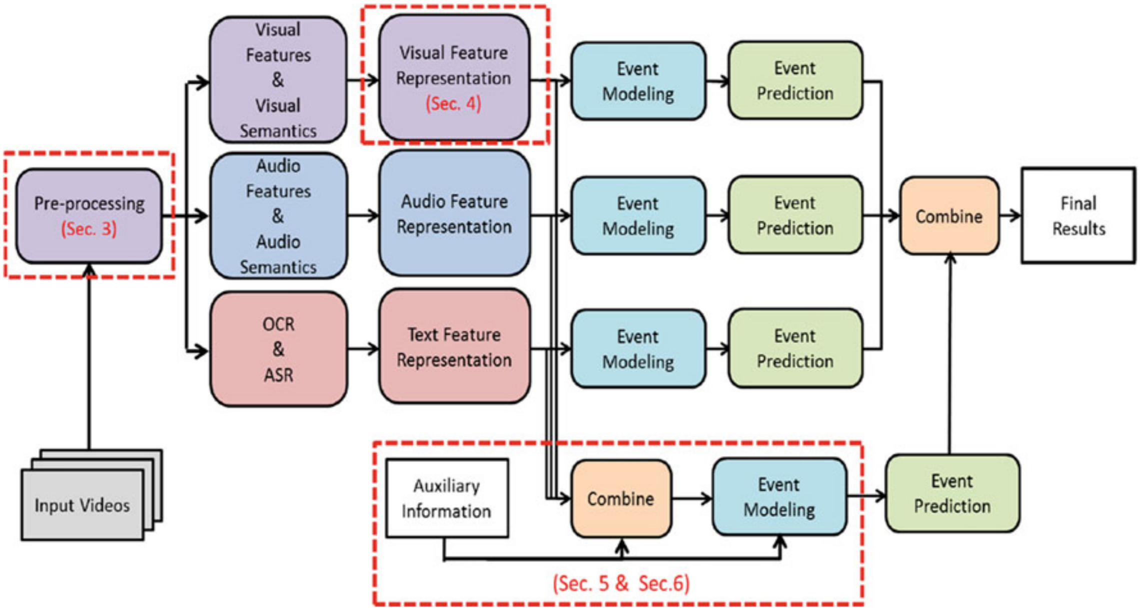


Fig. 2 Overview of the framework. The corresponding novel components described in Sections 3,4,5 and 6 are marked by *red boxes*

tion. However, this can dramatically improve the efficiency of feature extraction which could be very important for processing large scale video datasets.

For the feature representation, the spatial bag-of-words model achieves good performance by extending the classic bag-of-word model with spatial layout information. However, the commonly used spatial layout is very arbitrary and may not be effective to capture the complex spatial information embedded in different videos. In this paper, we systematically test a large number of different spatial layouts, i.e., tilings, to select the best one for each feature and each event. Our results show that the selected tilings can capture the spatial information more effectively than the commonly used tilings, thus improve event detection performance.

To jointly explore the multiple modalities provided by the features and the associated representations, we have developed an algorithm which learns a more robust and discriminative intermediate representation from multiple features so that better event models can be built. Finally, to tackle the more challenging problem of event detection with few positive exemplars, an innovative algorithm is developed which is able to effectively adapt the knowledge learnt from auxiliary sources to assist in event detection.

The rest of the paper is organized as follows. In Sect. 3, we introduce our studies on how video resolution can effect the efficiency and the effectiveness in event detection. In Sect. 4, we introduce our work on selecting the better tiling so that the spatial layout information can be encoded more effectively, and thus improve the performance in event detection. In Sects. 5 and 6, we introduce our novel methods developed to model the events with respect to different number of positive examples and the last section is our conclusion.

2 Related work

The study of multimedia event detection first emerged in structured scenarios, e.g., surveillance videos, sports and news videos. For example, in [1], a robust real-time detection method using multiple fixed-location monitors was introduced to detect unusual events in surveillance videos. In [48], an unsupervised online algorithm is developed for detecting unusual events in surveillance videos via dynamic sparse coding. Sadlier and O'Connor [33] proposed to use audio-visual features and support vector machine to detect events in field sports videos. Xu et al. [44] presented a novel approach for event detection from the live sports game using web-casting text and broadcast videos. Wang et al. [41] developed an multi-resolution bootstrapping framework framework for concept detection in news videos by exploring knowledge of sub-domain.

With the success of event detection in those structured videos, people started to focus more on the general unconstrained videos such as those obtained from internet video sharing web sites like YouTube. Since 2010, the ACM Multimedia society has launched three international workshops on events in multimedia and a couple of interesting works have been reported [28]. For example, Makkonen et al. [26] try to detect events by clustering videos from large media databases. In [4], build on recent work on local feature trajectories, the authors investigate the impact of a new trajectory filtering scheme and two new trajectory descriptors to the detection of such video events. Mertens et al. [27] exploit non-speech audio features for building acoustic super-models that detect complex events from low-level audio features, and show that even using audio alone they can achieve high recognition rates.

To facilitate and encourage the research of new technologies and algorithms for multimedia event detection, the TREC Video Retrieval Evaluation (TRECVID) workshop supported by National Institute of Standards and Technology (NIST) launched the task of "Event detection in Internet multimedia (MED)" in 2010 [31] which is often referred to as pre-specified multimedia event detection. In the MED task, a participant needs to detect the occurrence of an event within a video clip in the archive based on an event kit. According to the definition from NIST [31], the event kit defines an event which consists of:

- "An event name which is an mnemonic title for the event
- An event definition which is a textual definition of the event
- An evidential description is a textual listing of attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence indicating the event's existence but it is not an exhaustive list nor is it to be interpreted as required evidence
- A set of illustrative video examples each containing an instance of the event. The examples are illustrative in the sense that they help to form the definition of the event but they do not demonstrate all possible variability or potential realizations".

Since launched, the TRECVID MED task has quickly attracted many top research groups in both academic and industry and the TRECVID MED datasets have become the popular testing bed for multimedia event detection. Among the available datasets, the MED'11 training dataset is often used because of its moderate size and complexity. More specifically, there are 9,746 videos in the MED'11 training dataset which are about 300 h and belong to 18 events. Among the 18 events 3 events, e.g., P001 to P003 are from the MED'10 and the rest 15 events E001 to E015 are newly defined in MED'11. Table 1 list the name of the 18 events. There are totally 1,543 positive videos for the 18 events and the rest videos are background videos which do not belong to any of the 18 events.

To evaluate the performance in event detection, a couple of metrics are adopted by NIST for TRECVID MED tasks. Among them, the minimum normalized detection cost (minNDC) was used in the TRECVID MED'11 evaluation. The normalized detection cost (NDC) is computed as:

$$\mathrm{NDC} = \frac{C_{\mathrm{M}} \times P_{\mathrm{M}} \times P_{\mathrm{T}} + C_{\mathrm{FA}} \times P_{\mathrm{FA}} * (1 - P_{\mathrm{T}})}{\min(C_{\mathrm{M}} \times P_{\mathrm{T}}, C_{\mathrm{M}} \times (1 - P_{\mathrm{T}}))},$$

where P_{M} is the missed detection probability and P_{FA} is the false positive probability for the system of a given event. C_{M}, C_{FA} and P_{T} are predefined constants which are $C_{\mathrm{M}} = 80$, $C_{\mathrm{FA}} = 1$ and $P_{\mathrm{T}} = 0.001$, respectively. The minNDC is the minimum NDC a system can achieve on an event and the smaller value of minNDC indicates better performance.

The main observation from recent successful MED systems [2,9,14,17,20,29,31,32,39] is that the following components are in general important in achieving good performance in multimedia event detection:

The first important component is that a comprehensive set of features are required to be extracted from both video and audio channels so that different aspects of the information conveyed in the videos can be captured. Table 2 lists the features used by the E-LAMP system.

Second, those extracted raw features need to be converted to appropriate representations which can be utilized in modeling the multimedia event. The most widely used non-parametric representations are the bag-of-words model (BoW) [10] and its extension spatial bag-of-words model [19] which incorporates the spatial layout information in to the bag-of-words representation. For the parametric representation, the Gaussian Mixture Model [15] is the classical one and also shows good performance in multimedia event detection.

The third important component is how to model an event. Typically, the classical classification scheme is employed to model and detect the events when a relatively large number of positive training examples of an event are available. For example, the support vector machine with χ^2 kernel has shown good performance. A more challenging situation in

Table 1 Name of the events in MED'11 training dataset

E001	Attempting a board trick	E010	Grooming an animal
E002	Feeding an animal	E011	Making a sandwich
E003	Landing a fish	E012	Parade
E004	Wedding ceremony	E013	Parkour
E005	Working on a woodworking project	E014	Repairing an appliance
E006	Birthday party	E015	Working on a sewing project
E007	Changing a vehicle tire	P001	Making a cake
E008	Flash mob gathering	P002	Batting a run in
E009	Getting a vehicle unstuck	P003	Assembling a shelter

Table 2 Features used by the E-LAMP multimedia event detection system

	Visual feature	Audio feature
Low-level	SIFT [22], Color SIFT(CSIFT) [34], Motion SIFT(MoSIFT) [8], Transformed Color Histogram (TCH) [13], STIP [43], Dense Trajectory [42]	MFCC [49], Acoustic Unit Descriptors(AUDs) [7]
High-level	Semantic Indexing Concepts(SIN) [31], Object Bank [21], Optical Character Recognition (OCR) [20]	Acoustic Scene Analysis [6], Automatic Speech Recognition (ASR) [20]

modeling an multimedia event is when there are only very few positive examples because in practise precisely labeled training data are difficult to obtain. In this situation, the traditional classification scheme which works well using relatively large number of positive training examples might not be suitable anymore. Ma et al. [25] present a pioneer work to tackle this challenge using the knowledge adaptation.

The fourth component is how to fuse the multiple modalities to achieve good detection performance. The multiple modalities can come from different sources. For example, different features, same feature with different representations, different models built from different features, etc. Many fusion methods [3,12,30,35] have been proposed and in general they can be categorized into early fusion which fuses the feature representations or late fusion schemes which fuses the detection scores [37]. Recently, in [18] the authors propose a double fusion method which combines the early fusion and late fusion together so that the overall performance can be further improved.

Table 3 Time spent on extracting the MoSIFT feature over the TRECVID MED'11 training data which has about 300 h of videos

	Original video (h)	Resized video (h)
Resize	n/a	85
MoSIFT extraction	12,600	2,350
code book generation	1,800	1,800
Bag-of-words generation	1,800	930
Total	16,200	4,920

Table 4 Statistics of the video width from the MED'11 training dataset

Video width	Number of videos
640	3,506
320	2,000
128	1,336
540	1,333
480	635
Other	1,011

3 Improve efficiency of feature extraction by reducing video resolution

It has been shown that extracting a comprehensive set of features from the videos is an effective way to achieve good performance in multimedia event detection [2,9,14,17,20,29,31,32,39]. However, the feature extraction is in general computationally expensive and time consuming, especially for those motion features, e.g., MoSIFT, STIP and Dense Trajectory. This problem becomes more serious when the total hours of videos to be processed is large. For example, the TRECVID MED'12 testing dataset contains about 4,000 h of videos. In Table 3, we demonstrate this problem by showing the time spent on extracting MoSIFT feature over the MED'11 training dataset which is introduced in Sect. 2. There are 9,746 videos in this dataset and the total length of the video is about 300 h. From the table, we can see that it will take about 16,200 h to generate the MoSIFT feature using a common single CPU core which is more than 50 times realtime of the videos. To improve the efficiency of the feature extraction, we take a simple strategy which is to reduce the resolution of videos and then extract those motion features on the resized videos. More specifically, we reduce the resolution of videos according to the following criteria:

- If the width of the video is greater than 320 pixels, resize the video width to 320 pixels. The height of the video is resized according to the aspect ratio of the video.
- Otherwise, skip this video.

There are two reasons why we use 320 pixels as the target video width. The first ones is that our experiments show that this resolution preserves the vast majority of the features and further reducing the resolution significantly degrades the performance. The second reason is that a relatively large portion of the videos in the dataset will be resized. Table 4 shows the statistics of video resolution in the TRECVID MED'11 training dataset.

In Table 3, we show the time spent on video resizing and the extraction of MoSIFT feature over the resized videos. Compared to the raw videos, the total time spent on extracting MoSIFT features is reduced from 16,200 to 4,920 h

Table 5 Performance of MoSIFT feature extracted from the resized videos vs. original videos

	Original MoSift	Resized MoSift
E001	0.9004	**0.8996**
E002	0.9981	**0.9532**
E003	**0.5570**	0.6432
E004	0.5596	**0.5499**
E005	**0.7787**	0.7859
E006	0.9819	**0.9245**
E007	1.0019	**0.8936**
E008	0.5848	**0.5083**
E009	0.7460	**0.7458**
E010	**0.9134**	0.9204
E011	**0.9591**	0.9637
E012	**0.8694**	0.9151
E013	0.8067	**0.7097**
E014	**0.7140**	0.7336
E015	0.7994	**0.7278**
P001	0.6912	**0.6408**
P002	**0.4823**	0.5070
P003	0.8486	**0.8173**
Average	0.7885	**0.7689**

Bold values represent the best results

which is about three times faster. It clearly shows that on the resized videos the time spent on MoSIFT feature extraction is significantly less than that on the original videos. Furthermore, because fewer features are extracted on low resolution videos, the time spend on generating the bag-of-words model is also reduced dramatically.

We test the performance of the MoSIFT feature extracted from the resized videos on the TRECVID MED'11 training dataset. More specifically, we randomly sample half of the positive videos and null videos to form the training set and the rest half are used as testing set. To evaluate the performance, we adopt the minNDC as our evaluation metric which is introduced in Sect. 2. In general, smaller minNDC value represents better performance. Table 5 shows the performance of event detection using MoSIFT features extracted from resized videos vs. original videos. From the results, we observe that on average the performance of the resized MoSift is even a little bit better than that of the original MoSift. However, a simple *t* test shows that at 95 % significant level, the difference between the results from two methods is not statistically significant. For other motion features used in our system, e.g., STIP and Dense Trajectory, similar efficiency results to MoSIFT can be obtained but we omit them here.

4 Tiling

The spatial bag-of-words model (Spatial BoW) is the most widely used representations for raw features which is an extension of the classic bag-of-words model by incorporating the spatial layout information. In the spatial BoW model, an image is geometrically partitioned into several grids or tiles. A sperate histogram is then generated to describe each tile and the whole image is finally described as the concatenation of the histograms of all tiles. One problem of spatial bag-of-words model is that the existing tilings, e.g., the spatial partition of the image, are very limited which may not able to capture the versatile spatial information from the videos. For example in the spatial pyramid matching, the 1 × 1, 2 × 2 and 4 × 4 tiling are used and in [5] the 1 × 1, 2 × 2 and 1 × 3 tilings (shown in the first row in Fig. 3) are adopted. However, the use of those tilings is ad-hoc and some preliminary work has shown that other tilings might produce better performance [40]. To find more representative ways to encode the spatial information, we systematically tested about 80 different tilings to select the best one for each feature and each event. More specifically, our candidate tilings consitst of individual and the combination of the basic tilings which are: 1 × 1, 2 × 2, 3 × 3, 4 × 4, 1 × 2, 1 × 3, 1 × 4, 2 × 1, 3 × 1 and 4 × 1 tilings.

Table 6 shows the performance of feature specific tiling vs. our standard tiling, e.g., 1 × 1, 2 × 2 and 1 × 3 tilings, on MED'11 training dataset introduced in the Sect. 2. From the table, we can see clearly that for all of the five features, the feature specific tiling performs consistently better than the standard tiling.

Table 6 Performance of feature-specific tiling vs. standard tiling evaluated on the MED'11 training dataset using minNDC

Feature	SIFT	CSIFT	TCH	STIP	MOSIFT
Feature-specific tiling	0.6881	0.7006	0.7683	0.7854	0.6812
Standard tiling	0.6990	0.7262	0.7823	0.8005	0.7036

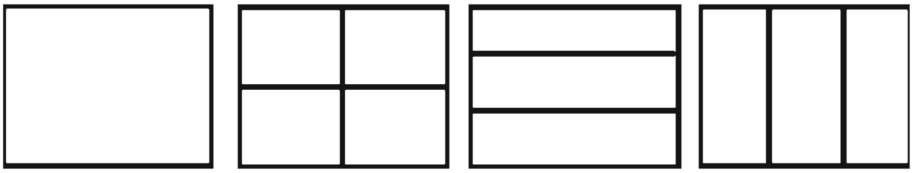


Fig. 3 Examples of tilings. From *left* to *right* 1 × 1, 2 × 2, 3 × 1 and 1 × 3 tilings, respectively

Table 7 Performance of event-specific tiling vs. standard tiling on Event E025 "Marriage Proposal"

Feature	SIFT	CSIFT	TCH	STIP	MOSIFT
Event-specific tiling	0.9486	0.9482	0.9562	0.8847	0.8823
Standard tiling	0.9712	0.9509	0.9654	0.8912	0.9712

Table 7 shows an example of the performance of event-specific tiling vs. classic tilings on a difficult event identified from MED'12 training data which is the event E025 "Marriage Proposal". It can be seen clearly that the event specific tiling can improve the performance over standard tiling.

5 A robust and discriminative intermediate video representation with sufficient training data

As we introduced in Sect. 2, how to jointly explore the multiple modalities is very important to achieve good performance. In general, the early fusion or late fusion schemes are often employed. In the early fusion, the feature vectors are concatenated or the kernel matrices computed from different features are combined, while in the late fusion the detection scores from the classifiers build upon different features are fused. In this section, we introduce an algorithm which is different from the early and late fusion schemes. The new algorithm learns an intermediate representation of videos from multiple features by exploiting both of the *target videos* and *external video* archives [25]. The target videos are the videos depicting the event to be detected. The external videos are the auxiliary labeled video archives that are used to help to learn the intermediate representation. The intermediate representation is a compact vector representation derived from multiple bag-of-words features of the videos through a transformation, during which the discriminative information is encoded. Meanwhile, our algorithm integrates representation inference and classifier training into a joint framework. In this way, the intermediate representation is tightly coupled with the loss function used by the classifier.

Compared with the original low-level feature representations, the learned intermediate representation is more robust and informative than the simple fusion of existing features and, therefore, better performance can be achieved [25]. In addition, the learnt intermediate representation is accurate to reflect the semantics which may help to bridge the semantic gap between the low-level features of a video and the semantic meaning of an event. Finally, since a robust loss function is used in our objective function, the performance is more reliable, when compared with other classifiers such as SVM.

Assume we have $X = [x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{n+m}]$ as the data matrix including the positive and negative examples $x_1, x_2, \dots, x_n$ of a particular event together with the external videos $x_{n+1}, \dots, x_{n+m}$. $Y = [y_1, y_2, \dots, y_n, y_{n+1}, \dots, y_{n+m}]$ indicate their labels. Note that the external videos have c classes and the positive and negative examples for an event are treated as two classes so we have $c + 2$ classes in total. The formulation of the proposed method can be illustrated as follows:

$$\begin{aligned} \min_{W,\Theta} \|X\Theta W - Y\|_{2,p} + \alpha \|W\|_F^2 \, . \\ s.t. \Theta^{\mathrm{T}}\Theta = I \end{aligned} \tag{1}$$

where Θ is the mapping function from video features to the intermediate representation and W is the classification matrix.

Note that our method is able to learn an intermediate representation coupled with the specific loss function. When the loss function changes, the intermediate representation, i.e., Θ changes accordingly.

The MED'11 training dataset introduced in the Sect. 2 and the development set from TRECVID 2011 semantic indexing task [5] are used to evaluate the algorithms. For the MED'11 training dataset, the videos from the MED'11 events are used as the *target videos* set while the *external video* set consists of the videos from 3 MED'10 events and the videos from the development set of TRECVID 2011 semantic indexing task. The training data comprise three parts. The first part consists of 100 positive examples and 500 negative examples randomly selected from the *target videos*. The second part includes the positive examples of MED 2010 from *external video* set. The third part is the data from the videos of the semantic indexing task. The remaining videos in the *target videos* set are our testing data. For all the videos, three features are extracted which are SIFT, CSIFT and MoSIFT. The raw features are first represented by the BoW model with the code book size of 4,096 respectively and then the BoW representations of the three features are further concatenated together as the final representation of each video. Detailed experimental setting can be found in [25]

Table 8 shows the results of our method vs. SVM using a χ^2 kernel. The evaluation metric used is minNDC [25].

It can be seen that our method outperforms the SVM classifier in most of the 15 event and the averages results over the 15 event is considerably better than that of SVM classifier. We also compared our method to a couple of other algorithms for example, AdaBoost, Linear Discriminant Analysis (LDA) followed by ridge regression, etc.; our method also outperforms those baselines considerably. For the detailed experimental results, please refer to [25].

Table 8 Performance of the proposed methods vs. SVM on MED'11 training dataset

Event name	SVM	SAIR
Attempting a board trick	0.826	**0.775**
Feeding an animal	**0.963**	0.964
Landing a fish	0.665	**0.626**
Wedding ceremony	0.466	**0.441**
Working on a woodworking project	0.726	**0.711**
Birthday party	0.885	**0.882**
Changing a vehicle tire	0.670	**0.636**
Flash mob gathering	0.629	**0.568**
Getting a vehicle unstuck	0.802	**0.711**
Grooming an animal	0.856	0.856
Making a sandwich	**0.821**	0.858
Parade	0.654	**0.632**
Parkour	0.570	**0.449**
Repairing an appliance	0.550	**0.508**
Working on a sewing project	0.706	**0.612**
Average	0.719	**0.682**

Bold values represent the best results

Table 9 Performance of SAR vs. SVM with different kernels on MED'11 training dataset

Kernel Type	SVM	SAR
RBF	0.954	**0.910**
χ^2	0.904	**0.881**

Bold values represent the best results

6 Structural adaptive regression with very few positive training examples

A more challenging problem in multimedia event detection is how to handle the situation when only very few positive training examples are available. This is because in practice precisely labeled training data are difficult to obtain. As a result, the traditional classification scheme which works well using relatively large numbers of positive training examples might not be suitable anymore. Recent research has shown that it can be beneficial to borrow knowledge from related tasks for multimedia analysis, especially when the number of training data is few [45,46]. Since there are some available video archives with annotated concept labels, we can leverage them to facilitate the multimedia event detection with only very few positive examples. Specifically, we propose to adapt the knowledge from concept level to assist the event modeling using the available video corpora with annotated concepts as our auxiliary resources. The difficulty is that the concepts from the auxiliary resources are different from the event to be detected. Hence, we have proposed a method to bridge the gap between the concepts and the event.

Denote the target training videos by $\tilde{X}_t = [\tilde{x}_t^1, \tilde{x}_t^2, \ldots, \tilde{x}_t^{n_t}]$. $y_t = [y_t^1, y_t^2, \ldots, y_t^{n_t}]^{\mathrm{T}}$ are the labels for the target training videos. $y_t^i = 1$ if the ith video x_t^i is a positive example whereas $y_t^i = 0$ otherwise. Denote the auxiliary videos by $\tilde{X}_a = [\tilde{x}_a^1, \tilde{x}_a^2, \ldots, \tilde{x}_a^{n_a}]$. $Y_a = [y_a^1, y_a^2, \ldots, y_a^{n_a}]^{\mathrm{T}}$ is their label matrix where c_a indicates that there are c_a different concepts. Y_a^{ij} denotes the jth datum of y_a^i and $Y_a^{ij} = 1$ if x_a^i belongs to the jth concept, while $Y_a^{ij} = 0$ otherwise. Then, our proposed formulation is:

$$\min_{W_a, W_t, b_a, b_t} \left\| \tilde{X}_a^{\mathrm{T}} W_a + 1_a b_a - Y_a \right\|_{2,1} + \left\| \tilde{X}_t^{\mathrm{T}} W_t + 1_t b_t \right. \\ \left. - y_t \right\|_{2,1} + \alpha \, \|W\|_{2,p} + \beta (\|W_a\|_F^2 + \|W_t\|_F^2) \qquad (2)$$

where $W = [W_a, W_t]$ and W_t is used for event detection in the target. The objective function can be optimized via an alternating approach described in [24]. We name the algorithm as structural adaptive regression (SAR).

We test our algorithm on the MED'11 training dataset as before and the development set from TRECVID 2011 semantic indexing task is used as the auxiliary videos set. In our experiments, all the videos are represented by SIFT and CSIFT BoW features. For the MED'11 dataset, we randomly sample ten positive and negative videos as the training set and the rest of the videos are used as the testing set. The experiments are independently repeated five times with randomly selected positive and negative examples. The detailed setting can be found in [24].

Table 9 shows the detection results of different approaches reported in [24]. We can see that our method SAR outperforms the SVM considerably. This indicates that it is beneficial to leverage auxiliary knowledge for event detection when we do not have sufficient positive examples. Another observation is that χ^2 kernel is better than RBF kernel for both SAR and SVM.

7 Conclusion

In this paper, we introduce a couple of cutting edge ideas to deal with both the effectiveness and efficiency challenges in multimedia event detection. More specifically, we have shown that reducing the resolution of raw videos to certain degree can dramatically improve the efficiency in feature extraction while not sacrificing the detection performance. We also discover that the standard tilings adopted by the spatial bag-of-words or spatial pyramid matching might not be able to effectively capture the spatial layout information of an event in videos; therefore, we suggest that more versatile tilings should be adopted. Finally, we introduce two event modeling methods which handle different situations with respect to different number of positive examples. In the future, we plan to focus more on another big unsolved problem which is that the semantic gap between the event model built by the system and the text event description is quite large. In other words, the event model built by the system is not sematic meaningful, thus difficult for human to understand and to interpret why a video detected by the system is a specific event. Event though we have incorporated some semantic features in event detection, e.g., ASR, OCR and SIN features, they are very noisy and inaccurate, thus not able to provide robust semantics [16]. Furthermore, those semantic features are utilized in a similar way as other low-level features in which they essentially just provide another type of "bag-of-words" representation of the video as other low-level features. To this end, their semantic meanings are not important anymore for modeling an event. To reduce the sematic gap, we believe that a possible solution is to first understand what semantic concepts are meaningful and discriminative in modeling an event and then find the mapping from the text description of the event into the available semantic visual concepts, spoken words on ASR OCR transcripts. Our preliminary work on this idea in the context of event detections with zero positive training example but only text description shows promising results [47].

Acknowledgments This work is supported in part by the National Science Foundation under Grant IIS-0917072 and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The US Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation 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 US Government.

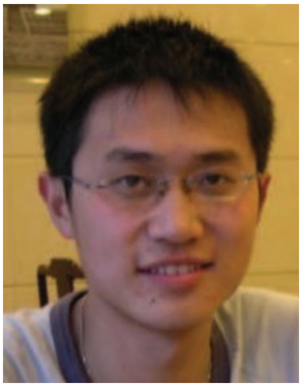
References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 555–560 (2008)
2. Akbacak, M., Bolles, R.C., Burns, J.B., Eliot, M., Heller, A., Herson, J.A., Myers, G.K., Nallapati, R., Pancoast, S., Hout, J.V., Yeh, E., Habibian, A., Koelma, D.C., Li, Z., Mazloom, M., Pintea,S., van de Sande, K.E., Smeulders, A.W., Snoek, C.G., Lee, S.C., Revatia, R., Sharma, P., Sun, C., Trichet, R.: The 2012 sesame multimedia event detection (med) system. In: TRECVID (2012)
3. Ayache, S., Quénot, G., Gensel, J.: Classifier fusion for svm-based multimedia semantic indexing. In: Advances in Information Retrieval, pp. 494–504. Springer, Berlin (2007)
4. Ballas, N., Delezoide, B., Prêteux, F.: Trajectories based descriptor for dynamic events annotation. In: Proceedings of the 2011 Joint ACM Workshop on Modeling and Representing Events, pp. 13–18. ACM, New York (2011)
5. Bao, L., Zhang, L., Yu, S.I., zhong Lan, Z., Jiang, L., Overwijk, A., Jin, Q., Takahashi, S., Langner, B., Li, Y., Garbus, M., Florian Metze, S.B., Hauptmann, A.: Informedia @ trecvid2011. In: TRECVID (2011)
6. Brown, G.J.: Computational auditory scene analysis: a representational approach (1992)
7. Chaudhuri, S., Harvilla, M., Raj, B.: Unsupervised learning of acoustic unit descriptors for audio content representation and classification. In: Interspeech (2011)
8. Chen, M., Hauptmann, A.: Mosift: Recognizing human actions in surveillance videos. Techical report, Carnegie Mellon University (2009)
9. Cheng, H., Liu, J., Ali, S., Javed, O., Yu, Q., Tamrakar, A., Divakaran, A., Sawhney, H.S., Manmatha, R., Allan, J., Hauptmann, A., Shah, M., Bhattacharya, S., Dehghan, A., Friedland, G., Elizalde, B.M., Darrell, T., Witbrock, M., Curtis, J.: Sri-sarnoff aurora system at trecvid 2012 multimedia event detection and recounting. In: TRECVID (2012)
10. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, vol. 1(2004)
11. Lan, Z., Bao, L., Yu, S.I., Liu, W., Hauptmann, A.G.: Double fusion for multimedia event detection. In: MMM (2012)
12. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: IEEE 12th International Conference on Computer Vision, 2009, pp. 221–228. IEEE, New York (2009)
13. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local color invariants. In: CVIU (2009)
14. Hill, M., Hua, G., Natsev, A., Smith, J.R., Xie, L., Huang, B.,Merler, M., Ouyang, H., Zhou, M.: Ibm research trecvid-2010 video copy detection and multimedia event detection system. In: TRECVID (2010)
15. Inoue, N., Shinoda, K.: A fast map adaptation technique for gmm-supervector-based video semantic indexing systems. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 1357–1360. ACM, New York (2011)
16. Jiang, L., Hauptmann, A., Xiang, G.: Leveraging high-level and low-level features for multimedia event detection. In: ACM Multimedia (2012)
17. Jiang, Y.G., Zeng, X., Ye, G., Ellis, D., Chang, S.F.: Columbia-ucftrecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In: TRECVID (2010)
18. Lan, Z.Z., Bao, L., Yu, S.I., Liu, W., Hauptmann, A.G.: Multimedia classification and event detection using double fusion. In: Multimedia Tools and Applications pp. 1–15 (2013)
19. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, vol. 2, pp. 2169–2178. IEEE, New York (2006)
20. Li, H., Bao, L., Gao, Z., Overwijk, A., Liu, W., fei Zhang, L., Yu, S.I., yu Chen, M., Metze, F., Hauptmann, A.: Informedia @ trecvid2010. In: TRECVID (2010)
21. Li, L.J., Su, H., Xing, E.P., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. Adv. Neural Inf. Process. Syst. **24** (2010)
22. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
23. Luo, J., Yu, J., Joshi, D., Hao, W.: Event recognition: viewing the world with a third eye. In: ACM Multimedia (2008)
24. Ma, Z., Yang, Y., Cai, Y., Sebe, N., Hauptmann, A.: Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In: ACM MM (2012)
25. Ma, Z., Yang, Y., Sebe, N., Hauptmann, A.: Multimedia event detection using a classifier-specific intermediate representation. IEEE Trans. Multimedia (2013)
26. Makkonen, J., Kerminen, R., Curcio, I.D., Mate, S., Visa, A.: Detecting events by clustering videos from large media databases.

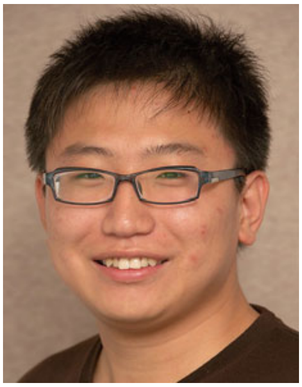
In: Proceedings of the 2nd ACM International Workshop on Events in Multimedia, pp. 9–14. ACM, New York (2010)
27. Mertens, R., Lei, H., Gottlieb, L., Friedland, G., Divakaran, A.: Acoustic super models for large scale video event detection. In: Proceedings of the 2011 Joint ACM Workshop on Modeling and Representing events, pp. 19–24. ACM, New York (2011)
28. Mezaris, V., Scherp, A., Jain, R., Kankanhalli, M., Zhou, H., Zhang, J., Wang, L., Zhang, Z.: Modeling and representing events in multimedia. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 613–614. ACM, New York (2011)
29. Natarajan, P., Natarajan, P., Manohar, V., Wu, S., Tsakalidis, S., Vitaladevuni, S.N., Zhuang, X., Prasad, R.: Bbn viser trecvid 2011 multimedia event detection system. In: TRECVID (2011)
30. Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R.: Multimodal feature fusion for robust event detection in web videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1298–1305. IEEE, New York (2012)
31. Over, P., et al.: Trecvid 2010—an introduction to the goals, tasks, data, evaluation mechanisms, and metrics. In: TRECVID (2010)
32. Perera, A., Oh, S., Leotta, M., Kim, I., Byun, B., Lee, C.,McCloskey, S., Liu, J., Miller, B., Huang, Z., Vahdat, A., Yang, W., Mori, G., Tang, K., Koller, D., Fei-Fei, L., Li, K., Chen, G., Corso, J., Fu, Y., Srihari, R.: Genie trecvid 2011 multimedia event detection: late-fusion approaches to combine multiple audio-visual features. In: TRECVID (2011)
33. Sadlier, D.A., O'Connor, N.E.: Event detection in field sports video using audio-visual features and a support vector machine. IEEE Trans. Circuits Syst. Video Technol. **15**(10), 1225–1233 (2005)
34. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. TPAMI (2010)
35. Schölkopf, B., Smola, A.J.: Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond. The MIT Press, Cambridge (2002)
36. Shyu, M.L., Xie, Z., Chen, M., Chen, S.C.: Video semantic event/concept detection using a subspace-based multimedia data mining framework. Trans. Multimedia (2008)
37. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402. ACM, New York (2005)
38. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: ACM Multimedia (2006)
39. Tamrakar, A., Ali, S., Yu, Q., Liu, J., Javed, O., Divakaran, A., Cheng, H., Sawhney, H.: Evaluation of low-level features and their combinations for complex event detection in open source videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 3681–3688. IEEE, New York (2012)
40. Viitaniemi, V., Laaksonen, J.: Spatial extensions to bag of visual words. In: ACM CIVR (2009)
41. Wang, G., Chua, T.S., Zhao, M.: Exploring knowledge of subdomain in a multi-resolution bootstrapping framework for concept detection in news video. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 249–258. ACM, New York (2008)
42. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR (2011)
43. Willems, G., Tuytelaars, T., Gool, L.V.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV (2008)
44. Xu, C., Wang, J., Wan, K., Li, Y., Duan, L.: Live sports event detection based on broadcast video and web-casting text. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 221–230. ACM, New York (2006)
45. Yang, J., Tong, W., Hauptmann, A.: A framework for classifier adaptation for large-scale multimedia data. Proc. IEEE (2012)
46. Yang, Y., Ma, Z., Hauptmann, A.G., Sebe., N.: Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans. Multimedia (2013)
47. Younessian, E., Quinn, M., Mitamura, T., Hauptmann, A.: Multimedia event detection using visual concept signatures. In: SPIE (2013)
48. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3313–3320. IEEE, New York (2011)
49. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of mfcc. J. Comput. Sci. Technol. (2001)

Author Biographies

Wei Tong is a postdoctoral researcher in the Language Technologies Institute at Carnegie Mellon University. He received his Ph.D. from the Department of Computer Science and Engineering at Michigan State University in 2010. His research interests include large-scale image/video retrieval, annotation and machine learning.



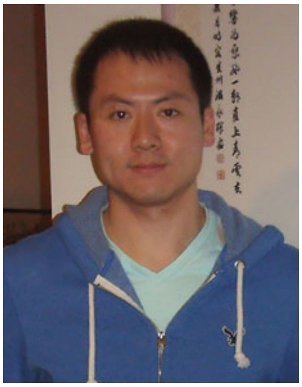
Yi Yang received the Ph.D degree in Computer Science from Zhejiang University, Hangzhou, China, in 2010. He was a postdoctoral research fellow at the University of Queensland from 2010 to May, 2011. After that, he joined Carnegie Mellon University. He is now a Postdoctoral Research Fellow at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include machine learning and its applications to multimedia content analysis and computer vision, e.g. multimedia indexing and retrieval, image annotation, video semantics understanding, etc.



Lu Jiang received his M.Sc. degree in Computer Science in 2011 and B.Sc. degree in Software Engineering in 2008, both from Xi'an Jiaotong University. Currently, he is a Ph.D student at school of computer science, Carnegie Mellon University. His research is focused on multimedia retrieval, web mining and large-scale machine learning.

Shoou-I Yu received the B.S. in Computer Science and Information Engineering from National Taiwan University, Taiwan in 2009. He is now a Ph.D. student in Language Technologies Institute, Carnegie Mellon University. His research interests include computer vision and multimedia retrieval.

ZhenZhong Lan received the B.S. in software engineering and statistics from Sun Yat-sen University, China in 2010. He is now a Ph.D. student at Languate Technologies Institute, Carnegie Mellon University. His research interests include computer vision and multimedia retrieval.



Zhigang Ma received the B.S. and M.S. both from Zhejiang University, Hangzhou, China in 2004 and 2006 respectively, and is currently working toward the Ph.D. degree from the University of Trento, Trento, Italy. He had been an intern at the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA from September, 2011 to March, 2012. From September, 2012 to February, 2013, he worked as a research assistant at the School of Computer Science, National University of Singapore, Singapore. His research interests include machine learning and its application to computer vision and multimedia analysis.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from The Johns Hopkins University, Baltimore, MD, USA, in 1982, the "Diplom" in computer science from the Technische Universität Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, USA in 1991. He is a Principal Systems' Scientist in the CMU Computer Science Department and also a faculty member with CMU's Language Technologies Institute. His research combines the areas of multimedia analysis and retrieval, man–machine interfaces, language processing, and machine learning. He is currently leading the Informedia project which engages in understanding of videodata ranging from news to surveillance, Internet video for applications in general retrieval as well as healthcare.