Effective Semantic Classification of Consumer Events for Automatic Content Management

Wei Jiang
Columbia University, New York, NY
wjiang@ee.columbia.edu

Alexander C. Loui
Eastman Kodak Company, Rochester, NY
alexander.loui@kodak.com

ABSTRACT

We study semantic event classification in the consumer domain by incorporating cross-domain and within-domain learning. An event is defined as a set of photos and/or videos that are taken within a common period of time, and have similar visual appearance. Events are generated from unconstrained consumer photo and video collections, by an automatic content management system, e.g., an automatic albuming system. Such consumer events have the following characteristics: an event can contain both photos and videos; there usually exist noisy/erroneous images resulting from imperfect albuming; and event data taken by different users, although from the same semantic category, can have highly diverse visual content. To accommodate these characteristics, we develop a general two-step Event-Level Feature (ELF) learning framework that enables the use of external data sources by cross-domain learning and the use of region-level representations, to enhance classification. Specifically, in the first step an elementary-level feature is used to represent images and videos. Then in the second step an ELF is constructed on top of the elementary feature to model each event as a feature vector. Semantic event classifiers can be directly built based on the ELF. Various ELFs are generated from different types of elementary-level features by using both cross-domain and within-domain learning: cross-domain approaches use two sets of concept scores at both image and region level that are learned from two external data sources; within-domain approaches use low-level visual features at both image and region level. Different types of ELFs complement each other for improved classification. Experiments over a large real consumer data set confirm significant improvements, e.g., over 90% MAP gain compared to the previous semantic event classification method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.4.m [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Management**]: Database Applications

WSM'09, October 23, 2009, Beijing, China.

General Terms

Algorithms, Experimentation.

Keywords

Semantic event classification, event-level feature, cross-domain learning, region-level representation, concept space.

1. INTRODUCTION

In this paper we explore the important issue of *semantic event classification*, i.e., classifying events organized by an automatic concept management system (e.g., an automatic albuming system) into pre-defined semantic event categories. An event is defined as a set of photos and/or videos that are taken within a common period of time, and have similar visual appearance. Events are generated from unconstrained consumer photo and video collections. For example, an event can be composed by photos and videos taken by any user at the 2009 commencement of a university.

Event mining has been an active research area for multimedia data analysis, and [23, 24] give some extensive surveys. To accommodate unconstrained consumer photos and videos where the lack of structured tags/descriptions usually exists, one popular approach used by automatic albuming systems is to generate events by chronological order and by visual similarities [3, 15, 18]. That is, media data that are taken at similar time (generally available as meta information) and with similar visual appearance are grouped together as an event. Our work in this paper is built upon such an automatic albuming system. We want to classify the organized events into a set of pre-defined semantic categories that are interesting to consumers, such as "wedding" and "birthday". Fig. 1 illustrates the position of our work in an event-based multimedia processing system.

The semantic event classification task has several characteristics. First, we need to process photos and videos simultaneously that often both exist in real consumer collections. Second, the algorithm needs to accommodate errors resulting from automatic albuming systems. For example, in Fig. 2, a photo irrelevant to "parade" is mistakenly organized into a "parade" event. Finally, events taken by different users, although from the same semantic category, can have quite diverse visual appearance, e.g., as shown in Fig. 2, data from two "parade" events can look very different. In comparison, sometimes we do not have enough event data for robust learning, e.g., in Kodak's consumer event collection we experiment on, there are only 11 "parade" events for training. The small sample learning difficulty may be encountered. This drives us to solicit help from cross-domain

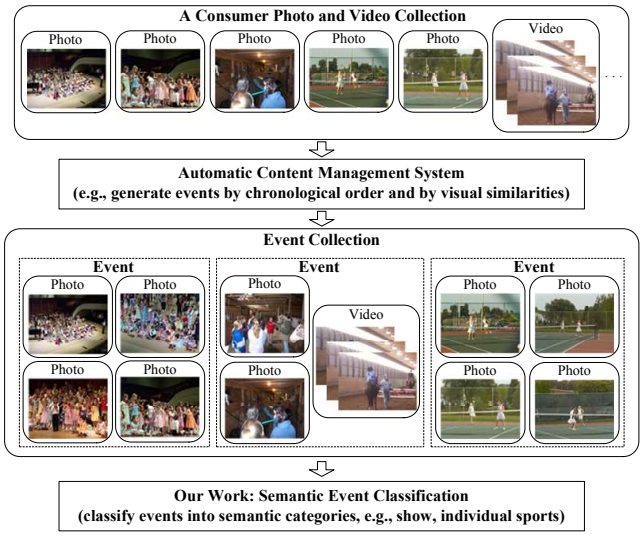

Figure 1: The position of our work in an event-based multimedia processing system.

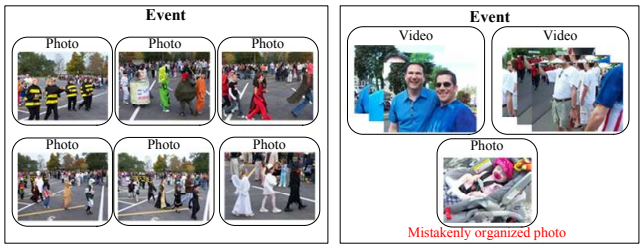

Figure 2: Two event data taken for different "parade" events, which have quite different visual content. These events are generated by an automatic albuming system, and in the event on the right a photo irrelevant to "parade" is mistakenly organized into this event.

learning [4, 12, 27] where we can borrow information from outside data sources such as TRECVID videos [21] or internet images to enhance our semantic event classification.

In our previous work [11], we have proposed an event classification algorithm to address the first two characteristics. An *Event-Level Feature* (*ELF*) representation is developed to model each event as a feature vector, based on which classifiers are directly built for semantic event classification. The ELF representation is flexible to accommodate both photos and videos at the same time, and is more robust to difficult or erroneous images from automatic albuming systems compared to the naive approach that uses image-level features to get classifiers straightforwardly.

In this paper, we systematically extend our previous work [11] to address all of the three characteristics. The contributions mainly lie in two folds.

(1) A general two-step ELF learning framework is proposed based on [11], as described in Fig. 3. In the first step each image (a photo or a video keyframe) is treated as a set of data points in an elementary-level feature space (e.g., a concept score space at the image level or a low-level visual space at the region level). Then in the second step a unified ELF learning procedure similar to [11] can be used to construct various ELFs based on different elementary features. Specifically an event-level *Bag-of-Features* (*BoF*) representation is developed to describe each event as a feature vector, which is directly used for classification.

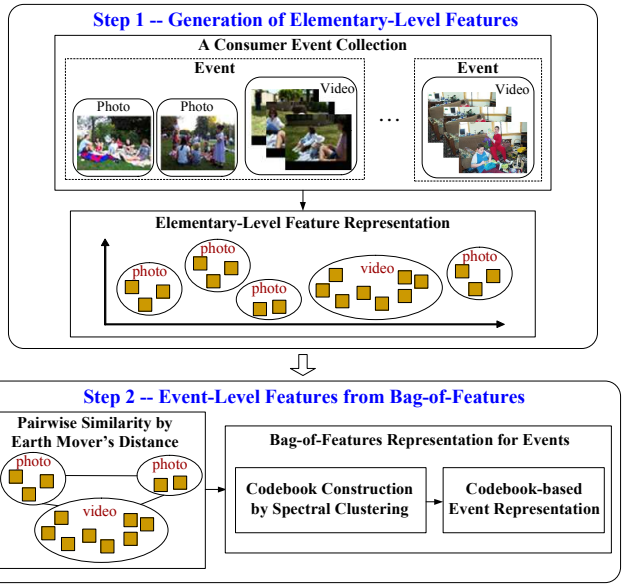

Figure 3: The general ELF learning framework. In the first step, each image (a photo or a video keyframe) is treated as a set of data points in an elementary-level feature space, and then in the second step, an ELF representation can be constructed with the learning process similar to [11].

(2) Using the general ELF learning framework, we conduct cross-domain and within-domain learning for semantic event classification, as described in Fig. 4. For cross-domain learning, we adopt the PRED framework [4]. That is, a set of models that are built based on the old data source are applied to the current data to generate predictions. Such predictions are then used as features to represent the current data and to learn models in the current domain. In practice our cross-domain approaches incorporate two sets of concept detection scores from pretrained models over two different old data sources, at both image and region level. Each set of concept scores forms an elementary-level concept space that is then used to construct a cross-domain ELF. These old data sources are: the TRECVID 2005 broadcast news video set [21] with a 374-concept LSCOM ontology [13]; and the LHI image-parsing ground-truth data set with a 247-concept regional ontology [25]. Within-domain approaches use low-level visual features over entire images or image region segments as elementary-level visual features to learn within-domain ELFs. The cross-domain and within-domain ELFs complement and cooperate with each other to achieve improved classification performance.

Compared with the previous method [11], this new approach has the following advantages:

(1) Both image-level and region-level features are used to construct ELF representations, while [11] only uses the global image-level feature. As have been shown by many previous works [10, 26], local regional features can complement global image-level features by capturing the detailed object information, and can greatly help classification.

(2) Both cross-domain and within-domain learning are incorporated at both image and region level to enhance classification. The general ELF learning framework

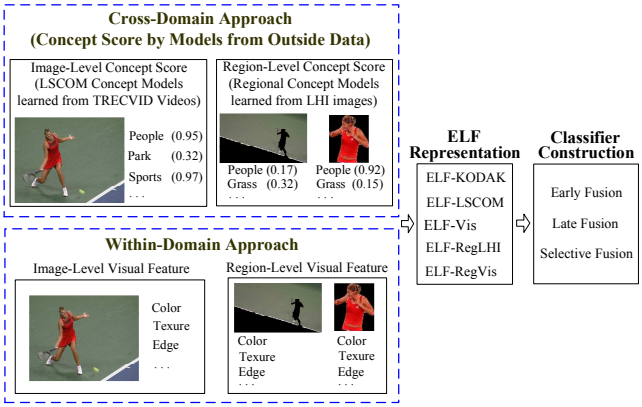

Figure 4: The overall framework of our approach. Compared with the previous method, we use both cross-domain and within-domain learning at both image and region level to enhance classification. In total of five ELFs are constructed which complement and cooperate with each other to get improved classification.

enables us to incorporate large external data collections as well as large ontologies to help semantic event classification. Large ontologies provide rich descriptors to represent the media content, and large external data collections import previously learned knowledge to help classification.

We evaluate our algorithm over Kodak's event data collection from real consumers, which contains event data from 10 semantic categories that are defined by real users as interesting. As will be shown in the experiments, significant performance gains can be achieved by combining different ELFs. For example, the overall MAP can be improved by more than 90% compared to the previous semantic event classification method.

2. RELATED WORKS

In this section we briefly review some related works about event mining and cross-domain learning.

2.1 Event Mining

Events can be defined as real-word occurrences that unfold over space and time. Event mining has been an active research area for multimedia analysis. In [24], a 5W1H framework is used to describe events, i.e., who?, when?, where?, what?, why?, and how?. Structured descriptive information and/or abundant metadata are required for such event description. In this work, we consider the task of classifying unconstrained consumer photos and videos, where the multimedia data usually do not have structured tags or descriptions, and the only meta information available in many cases is the taken date and time that comes along with the digital devices. To accommodate such unconstrained consumer collections, one popular approach for automatic albuming systems is to organize the multimedia data by chronological order and by visual similarities [3, 15, 18]. Our work in this paper is built upon such automatic albuming systems, i.e., to classify the organized events into semantic categories.

There are many previous works exploring event modeling and classification. One major branch is to detect object action from the continuous capture. For example, by extracting and tracking foreground objects (e.g, human body, vehicle, etc.), in [5, 19] action recognition is conducted based on object motion. By grammar-based modeling of tracked object parts, in [9, 16] known events are recognized by model-based matching. However, object extraction and tracking in unconstrained consumer videos is known to be extremely difficult, and we need to deal with both photos and videos that often exist in event data at the same time. Thus these object-action-detection methods can not be applied. Another important branch for event classification is to detect domain-specific events inherent in video sequences, e.g., to detect regular patterns in sports videos corresponding to recurrent semantic events like goal, kick, whistle [7, 8, 28]. However, due to the unconstrained content of consumer event data, there hardly exists effective domain knowledge or rule for our semantic event classification task.

2.2 Cross-Domain Learning

Cross-domain learning [4, 12, 27] is recently proposed as a potent technique to port information from outside data sources (domains) for helping analyze the current data. Such borrowed information can be selected data points from an outside domain or learned models over outside source data. Assume that we have a set of data $\mathcal{D}^o$ from an old domain (e.g., broadcast news videos from TV programs), which has been well studied, and now we want to examine a new data set $\mathcal{D}^c$ in the current domain (e.g., consumer event data). The model adaptation method, e.g., Adaptive SVM [27], tries to learn a new classifier by oscillating the old classifier built over the outside data source $\mathcal{D}^o$, so that the new classifier is close to the original old classifier and can separate the new current data $\mathcal{D}^c$. The data adaptation approach, e.g., Cross-Domain SVM [12], selectively uses the most important data from the outside domain to help build the new classifier. The importance of a data point from outside source $\mathcal{D}^o$ is determined by: the discriminative information it carries about separating the outside data $\mathcal{D}^o$; and the similarity between this point and the data distribution of the current domain $\mathcal{D}^c$. However, both the model adaptation and data adaptation methods have difficulties in applying to our semantic event classification task, due to mainly two reasons. First, our consumer event collection is very different from the outside data sources (e.g., TRECVID news videos or LHI images). The restriction of Adaptive SVM that new classifiers in the current domain have to be close to the old classifiers from the outside domain makes it difficult to capture the dramatic domain difference. Second, sometimes we only have a small set of event data for training, and if we incorporate a lot of data from the outside data source that has very different data distribution with the current event collection, the learned new classifier by Cross-Domain SVM can be biased by the outside data.

To avoid the above issues, in this paper we use another popular cross-domain learning approach, which can be described as follows. A set of models are built based on the old data $\mathcal{D}^o$ and are applied to the current data $\mathcal{D}^c$ to generate predictions. Such predictions are then used as features to represent $\mathcal{D}^c$ and to learn models in the current domain. This approach is usually called the "PRED" method [4] since predictions generated from $\mathcal{D}^o$ play the role of porting information to $\mathcal{D}^c$. This PRED approach is attractive due to its flexibility where no assumption about the underlying data distribution is needed.

3. THE ELF LEARNING PROCESS

We begin with the terminologies, following the notations used in [11]. Assume that we have a large data collection, including photos and videos from consumers. The entire data set is partitioned into a set of *macro-events*, and each macro-event is further partitioned into a set of *events*. The partition is based on the capture time of each photo/video and the color similarity between photos/videos, by using previously developed automatic albuming systems like [15]. Let E^t denote the t-th event which contains m_p^t photos and m_v^t videos, and I_i^t and V_j^t denote the i-th photo and j-th video in E^t, respectively. Our target is to classify E^t into a set of pre-defined semantic categories $S_{E1}, \ldots, S_{EL}$.

3.1 ELF Representation

In the previous work [11], we have proposed a BoF representation at the event level to describe each event as a feature vector, based on which semantic event classifiers can be directly built. Specifically, a set of 21 SVM concept detection models [1] that are learn based on Kodak's consumer benchmark video set (5166 keyframes from 1358 consumer videos) [14] are applied to the consumer event data, which generate concept detection scores for 21 consumer concepts. These 21 concepts are chosen to be important to consumers based on real user study [14]. The consumer videos for training the concept detectors and the consumer event data we study in [11] and in this work are from the same Kodak's consumer data pool, or the same domain. The resulting concept detection scores are used to represent the global event images, based on which an ELF is constructed through the BoF technique. Specifically, pairwise similarities between image/video data point sets are calculated using the Earth Mover's Distance [20] which can deal with different set sizes (i.e., different images/videos have different numbers of data points). Then a codebook is constructed through spectral clustering [17] over the pairwise similarity matrix, where each cluster corresponds to one codeword in the codebook. The codewords span a feature space in which each event E^t can be represented as a codebook-based feature vector. Using the ELF vector, SVM classifiers can be learned to conduct semantic event classification. As demonstrated in [11], the ELF representation alleviates the influence of difficult or erroneous images from automatic albuming systems in measuring event-level similarities. Superior performance can be obtained compared to the counterparts using straightforward image-level features. The second step in Fig. 3 illustrates the process of generating ELF representations.

3.2 The General ELF Learning Framework

We develop a general two-step ELF learning framework based on [11] as shown in Fig. 3. In the first step, elementary-level features are generated to represent each photo I_i^t or video V_j^t as a set of data points, e.g., I_i^t can be treated as a single-point set with an image-level low-level visual feature $\mathbf{f}(I_i^t)$, or a multipoint set with region-level low-level visual features $\{\mathbf{f}(r_{i1}^t), \ldots, \mathbf{f}(r_{in}^t)\}$ where each r_{ik}^t is a region from image I_i^t described by a feature vector $\mathbf{f}(r_{ik}^t)$. Then in the second step, based on different types of elementary-level features, the previous ELF learning process can be used to construct ELF representations.

Compared with the previous method in [11], this new approach has several advantages. First, both image-level and region-level features are used to construct ELF representations, while [11] only uses the global image-level feature. As have been shown by many previous works [10, 26], local regional features can complement global image-level features by capturing the detailed object information, and can greatly help classification. In addition, both cross-domain and within-domain learning are incorporated at both image and region level to enhance classification. The general ELF learning framework enables us to incorporate large external data collections as well as large ontologies to help semantic event classification. Large ontologies provide rich descriptors to represent the media content, and large external data collections import external knowledge to help classification.

4. SEMANTIC EVENT CLASSIFICATION WITH MULTITYPE ELFS

The above ELF learning framework is very flexible. Different types of elementary-level features can be used to generate ELFs. In this work, we construct ELFs by both cross-domain and within-domain learning. We adopt the PRED cross-domain learning technique where two sets of concept scores at both image and region level are obtained from two external data sources. They form two elementary-level concept spaces, based on which two sets of cross-domain ELFs are generated. For within-domain learning, low-level visual features at both image and region level are used as elementary features, on top of which within-domain ELFs are generated. SVM-based semantic event classifiers are learned over these ELFs, and through different fusion approaches, significant performance improvement can be achieved by selectively combining various ELFs. Fig. 4 summarizes the overall approach.

4.1 Cross-Domain ELFs

We further categorize the cross-domain ELFs as image-level or region-level, i.e., concept detectors from external data sets are learned at the image or region level to generate the image-level or region-level elementary concept spaces.

Image-level concept space – We use the TRECVID 2005 news video set [21] with a 374-concept LSCOM ontology [13] to provide the old-domain knowledge for generating a concept-score-based ELF at the image level. The LSCOM ontology [13] contains 449 multimedia concepts related to events, objects, locations, people, and programs. The entire TRECVID 2005 development set (61901 subshots) [21] is labeled to this ontology. By using visual features over the entire image, i.e., 5×5 grid-based color moments, Gabor texture, and edge direction histogram [2], a total of 374 SVM concept detectors are learned based on the labeled TRECVID subshots. These concepts are those with high-occurrence frequencies in LSCOM.

We apply the concept detectors learned from the external TRECVID data to obtain the concept detection probabilities for each image x (a photo or a video keyframe) in the current event data set. These probabilities represent x in a concept space with a feature vector formed by concept scores $\mathbf{f}_c(x) = [p(C_1|x), \ldots, p(C_m|x)]^T$, where each C_k is a concept. Each photo is a single-point set and each video is a multipoint set in the concept space. Then the ELF learning process described in the second step of Fig. 3 can be used to generate the ELF over the LSCOM ontology, which is called *ELF-LSCOM*.

Region-level concept space – Region-level features provide detailed object information to describe the image content, which is complementary to global image-level features. In the regional approach, each image x is segmented into a set of regions $r_1, \ldots, r_n$, and each region can be represented by a feature vector in either the concept space (this subsection) or the low-level visual space (Sec. 4.2). In the elementary region-level feature space, both photos and videos are treated as multipoint sets, and the ELF learning procedure from the second step of Fig. 3 can be conducted to obtain ELF representations.

To generate region-level concept scores, we need external region-level concept detectors. In this work, the LHI image-parsing ground-truth data set (the free version) [25] is used to build region-level concept detectors. The data set contains images from 6 categories: manmade object, natural object, object in scene, transportation, aerial image, and sport activity. These images are manually segmented and the regions are labeled to 247 concepts. Fig. 5 gives an example image and its manual region annotation. Low-level visual features, i.e., color moments, Gabor texture, and edge direction histogram, are extracted from each region. By using each region as one sample, SVM classifiers are trained to detect the 247 region-level concepts. These detectors generate concept detection scores for each automatically segmented region in our event data set. Then an ELF (*ELF-RegLHI*) can be learned based on the region-level concept scores.

Figure 5: An example image and its corresponding manual annotation from the LHI image set. Every segmented region in this image is annotated and in the figure we only show annotations for some large regions.

4.2 Within-Domain ELFs

The use of concept score space has been proved effective for semantic annotation by several previous works [6, 11]. However, low-level visual features are still indispensable for semantic event classification, especially when we only have a limited concept ontology. Since in practice we cannot train a concept detector for every possible concept in every aspect of our life, low-level visual features can capture useful information not covered by the available concept detectors. Similar to the case of cross-domain learning, within-domain visual-feature-based approaches can also be categorized as using image-level or region-level visual features.

With image-level visual features, each image x (a photo or a video keyframe) is represented as a low-level visual feature vector $\mathbf{f}_l(x) = [f_1(x), \ldots, f_d(x)]^T$. Then each photo is a single-point set and each video is a multipoint set, based on which an ELF (*ELF-Vis*) can be generated. Specifically, we use the same low-level visual features as the ones for getting image-level concept detection scores, i.e., grid-based color moments, Gabor texture, and edge direction histogram.

Using region-level visual features, each region is represented as a low-level visual feature, and the entire image is a multipoint set in the regional feature space (so is a video), based on which we generate an ELF (*ELF-RegVis*). In practice, we also use the same low-level visual features as the ones for getting region-level concept detection scores, i.e., color moments, Gabor texture, and edge direction histogram.

In addition to the above 4 types of ELFs, we keep the ELF representation learned from the previous work [11], where Kodak's consumer benchmark video set with the 21-concept consumer ontology [14] is used to train concept detectors and to generate concept detection scores as elementary-level features for constructing the ELF. We call this ELF representation *ELF-KODAK*. This can also be treated as a within-domain learning approach.

4.3 Classification with ELFs

By now we have five ELFs learned from different types of elementary-level features: ELF-KODAK, ELF-LSCOM, ELF-RegLHI, ELF-Vis, and ELF-RegVis. Individual classifiers can be built over each type of feature for semantic event classification, and improved performance can be expected if we appropriately fuse these ELFs. In early fusion, we concatenate these ELFs into a long feature vector based on which classifiers can be trained. In late fusion, we combine the classifiers individually trained over each ELF. Also, we can use selective fusion, i.e., forward feature selection in the manner of both early and late fusion. In selective early fusion, we gradually concatenate one more ELF at one time based on the cross-validation error rate to choose the optimal combination of features. Similarly, in selective late fusion, we gradually combine one more classifier trained over individual ELFs. In practice, the SVM classifier [22] with the RBF kernel is used as semantic event classifiers based on ELFs, since SVM has been proved effective by many previous literatures for image/video classification [1, 21].

5. EXPERIMENTS

We evaluate our algorithm over 1972 consumer events in a consumer event collection from Kodak. These events are generated from the automatic albuming system described in [15], and are labeled to 10 different semantic event categories. Table 1 gives the detailed definitions of these semantic event categories. Fig. 6 shows the example events for these semantic categories. More details about this event data set can be found in [11]. A total of 1261 events are randomly selected for training, and the rest are used for testing. The training/testing data are partitioned at the macro-event level, i.e., events from the same macro-event stay together for training/testing. This avoids the situation where similar events from the same macro-event are separated, which will simplify the classification problem.

The average precision (AP) and mean average precision (MAP) are used as performance measures, which are official metrics for video concept classification [21]. To calculate AP for a semantic event category, we first rank the test data according to the classification posteriors of this semantic category. Then from top to bottom, the precision after each positive sample is calculated. These precisions are averaged over the total number of positive samples for this semantic category. AP favors highly ranked positive samples and combines precision and recall values in a balanced way. MAP is calculated by averaging APs across all semantic event categories. Parameters in SVM are tuned by cross validation within the training set, where for the RBF kernel

Table 1: Definition of semantic event categories in Kodak's consumer event data set.

semantic event category	definition
wedding	bride, groom, decorated car, wedding cake or reception, bridal party, or events about the wedding day
birthday	birthday cake, birthday balloon, wrapped birthday presents, birthday caps
Christmas	Christmas tree and usual Christmas decoration, not necessarily taken on Christmas day
parade	processing of people or vehicles moving through a public place
picnic	outdoor, with or without a picnic table, with or without shelter, food and people in view
team sport	basket ball, football, baseball, hockey, and other team sports
individual sport	tennis, swimming, bowling, and other individual sports
animal	pets (e.g., dogs, cats, horses, fish, birds, hamsters), wild animals, zoos, and animal shows
school activity	school graduation, school days (first or last day of school), and other events related to school
show	show and concert, recitals, plays, and other show events

Figure 6: Examples of consumer event data for different semantic event categories.

$K(\mathbf{x},\mathbf{y}) = \exp\{-\gamma||\mathbf{x}-\mathbf{y}||_2^2\}$, $\gamma = (1/d)^{2^t}$, d is the dimension of data point $\mathbf{x}$, and t is chosen from $t = \{-3,-2,-1,0,1\}$; and the error control parameter C in SVM [22] is chosen from $C = 2^s$ and $s = \{0,1,2,3,4\}$. Table 2 gives some detailed information of different ELFs.

Table 2: Information of different ELFs.

	dimension	type
ELF-KODAK	97	Image level, within-domain
ELF-LSCOM	93	Image level, cross-domain
ELF-Vis	80	Image level, within-domain
ELF-RegVis	83	Region level, within-domain
ELF-RegLHI	80	Region level, cross-domain

Fig. 7 gives the AP performance for each semantic event category and the overall MAP using different individual ELFs. From the result, different types of ELFs have different advantages in classifying different semantic event categories. In general, image-level concept scores (ELF-KODAK and ELF-LSCOM) perform well over complex semantic event categories like "birthday", "Christmas", "parade", "picnic", "school activity", and "wedding", which are composed by many constructive concepts, e.g., wedding consists of wedding gowns, suits, park, flowers, etc. The concept scores capture the semantic information about occurrences of these constructive concepts, and are expected to be superior to low-level features in classifying such semantic events. On the other hand, ELF-Vis performs extremely well over semantic event categories that are determined by only one or a few concepts, such as "animal", where the detection scores for other constructive concepts are not so helpful. Similarly ELF-RegLHI performs well over complex semantic event categories in general, and it works very well over those semantic events having strong regional cues, e.g., "individual sport" or "show", where detection of sport fields and swimming pools, or stages and people in costume help a lot.

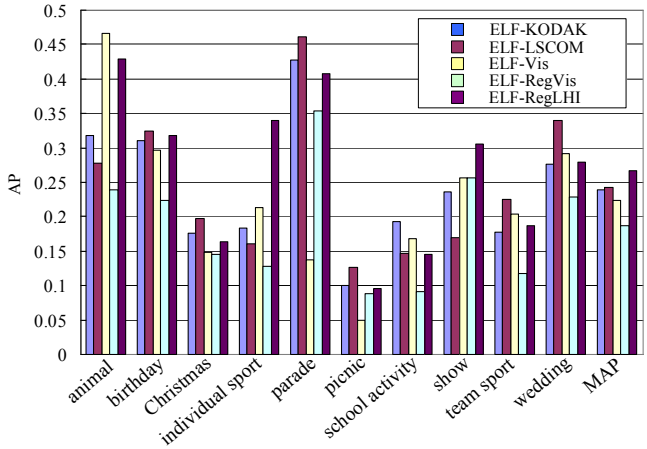

Figure 7: Performances of individual ELFs.

In terms of image-level concept scores, the large ontology (ELF-LSCOM) outperforms the small one (ELF-KODAK), although concept detectors for the later are trained with consumer videos that are more similar to our consumer event data than the TRECVID data. This confirms that a large ontology can provide rich descriptors to represent the media content and a large external data source (even with very different data distribution to our event collection) can be quite helpful. Specifically, ELF-LSCOM gets very good results over "parade", "team sport", and "wedding". This is because the TRECVID news videos and the LSCOM ontology provide good detectors for many constructive concepts related to parade (e.g., protest, demonstration, etc), sports (e.g., basketball, football, etc.), and well-suited people (e.g., corporate leader, government leader, and so on). Fig. 8 gives some example keyframes from the TRECVID news videos for some related constructive concepts. We can see that detection of such constructive concepts is very helpful.

Figure 8: Example keyframes from TRECVID news videos for some constructed concepts useful to help our semantic event classification. For example, the Government Leader concept detector captures well-suited people, which helps classify "wedding" events; the Parade concept detector trained over the large TRECVID set can provide useful information to help classify consumer "parade" events where only a few event data is available for training; the Sports concept detector captures different types of team sports like basketball and football, and can greatly help classify consumer "team sport" events.

Fig. 9 shows performances of different fusion methods, and the best individual ELF result is also given for comparison. From the result, consistent performance improvements can be achieved over every semantic event when we combine different ELFs by either early fusion or late fusion, i.e., about 35% gain in MAP compared to the best performing individual ELF. In addition, by selectively combining different types of ELFs, further performance gain can be obtained. Compared to the best individual ELF, the selective fusion approach can get more than 70% MAP improvement. Also, compared to the ELF-KODAK method (which is used in the previous work [11]), MAP is improved by more than 90%. Fig. 10 gives the top 10 events with highest classification scores for "birthday" and "show", using the previous ELF-KODAK [11] and the selective late fusion. Each event is marked by a rectangle, green as correct and red as incorrect. The results show clear performance improvements by using information from multiple types of ELFs.

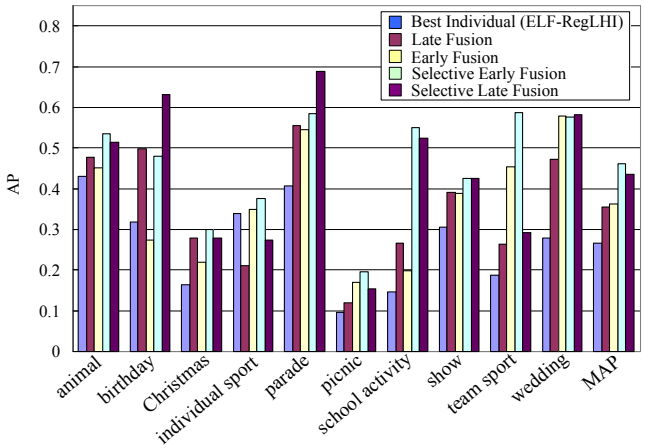

Figure 9: Performances of different fusion methods. Significant improvements can be achieved by selectively combining different ELFs.

Fig. 11 gives the details of which types of ELFs are actually used by the selective early fusion approach for classifying different semantic events. This figure clearly shows the contributions of various ELFs. From the figure, ELF-RegLHI is always chosen by all semantic event classifiers. This is consistent with the previous results (Fig. 7) that ELF-RegLHI is the best performing individual feature. ELF-RegVis is the least favorite feature and is only used by the "animal" classifier and the "team sport" classifier. This is reasonable since ELF-RegVis has bad classification performance over most semantic events, but can help classify simple ones like "animal" that are composed by one or a few objects. ELF-LSCOM helps with 4 semantic events that have high support from TRECVID news videos, i.e., good detectors can be obtained from TRECVID data to detect useful constructive concepts. This is also consistent with our previous analysis.

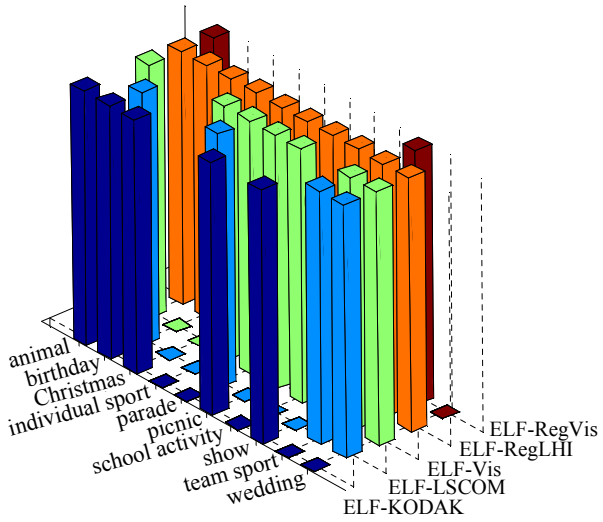

Figure 11: Detailed usage of different ELFs by semantic event classifiers with selective early fusion. ELF-RegLHI is the most popular feature. ELF-RegVis is the least favorite feature. ELF-LSCOM is used for semantic events with high support from TRECVID news videos where good detectors can be obtained from the TRECVID data to detect useful constructive concepts.

6. CONCLUSIONS

We develop a general two-step ELF learning framework for semantic event classification, which is flexible to incorporate various elementary-level features in the first step for learning different ELFs in the second step through a unified procedure. We adopt both cross-domain and within-domain learning to generate ELFs at both image and region level, based on both concept-score spaces and low-level visual spaces. Experiments over a real consumer event collection demonstrate significant performance improvements by combining different ELFs. Future work includes incorporating more types of elementary-level features and more external data sources/ontologies to generate ELFs, and further performance gain can be expected. In addition, the temporal information from the event boundaries can be used, and the relationship between the visual codebooks can be modeled for better event description.

7. REFERENCES

[1] S. Chang and et al. Large-scale multimodal semantic concept detection for consumer video. In *ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, pages 255–264, 2007.

(a) Top 10 ranked events for "birthday"

(b) Top 10 ranked events for "show"

Figure 10: Examples of top ranked events according to the classification posterior in descending order. Each event is marked by a rectangle, green as correct and red as incorrect. At most four images (photos or video keyframes) are shown for each event.

[2] S. Chang and et al. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.

[3] M. Cooper and et al. Temporal event clustering for digital photo collections. In *ACM Trans. Multimedia Comput. Commun.*, 1(3):269–288, 2005.

[4] H. Daumé III. Frustratingly easy domain adaptation. In *Annual Meeting of the Asso. of Comp. Linguistics*, 2007.

[5] J.W. Davis and A.F. Bobick. The representation and recognition of action using temporal templates. In *IEEE Int'l Conf. on CVPR*, pages 928–934, 1997.

[6] S. Ebadollahi and et al. Visual event detection using multi-dimensional concept dynamics. In *IEEE Int'l Conf. on Image & Expo*, pages 881–884, 2006.

[7] A. Ekin, A. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. In *IEEE Trans. Multimedia*, 12(7):796-807, 2003.

[8] Z. Feng and et al. Video data mining: Semantic indexing and event detection from the association perspective. In *IEEE Trans. Knowledge Data Engineering*, 17(5):665-677, 2005.

[9] Y.A. Ivanov and A.F. Bobick. Recognition of visual activities and interactions by stochastic parsing. In *IEEE Trans. PAMI*, 22(8):852-872, 2000.

[10] L.J. Li, R. Socher, and F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Int'l Conf. on CVPR*, Miami, FL, June, 2009.

[11] W. Jiang and A. Loui. Semantic event detection for consumer photo and video collections. In *ICME*, 2008.

[12] W. Jiang and et al. Cross-domain learning methods for high-Level visual concept classification. In *IEEE Int'l Conf. on Image Processing*, pages 161-164, 2008.

[13] LSCOM lexicon definitions and annotations v1.0. Columbia Univ. ADVENT Tech. Report, 2006.

[14] A.Loui and et al. Kodak consumer video benchmark data set: concept definition and annotation. In *ACM SIGMM Int'l Workshop on Multimedia Information Retrieval*, pages 245–254, 2007.

[15] A. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. In *IEEE Trans. Multimedia*, 5(3):390–402, 2003.

[16] G. Medioni and et al. Event detection and analysis from video streams. In *IEEE Trans. PAMI*, 23(8):873–889, 2001.

[17] A. Ng and et al. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[18] J.C. Platt, M. Czerwinski, and B. A. Field. PhotoTOC: Automatic clustering for browsing personal photographs. In *IEEE Pacific Rim Conf. Multimedia*, 2003.

[19] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. In *IJCV*, 50(2):203–226, 2002.

[20] Y. Rubner and et al. The earth mover's distance as a metric for image retrieval. In *IJCV*, 40(2):99–121, 2000.

[21] TRECVID http://www-nlpir.nist.gov/projects/trecvid

[22] V. Vapnik. *Statistical learning theory*. Wiley-Interscience, New York, 1998.

[23] U. Westermann and R. Jain. Toward a common event model for multimedia applications. In *IEEE MultiMedia archive*, 14(1):19–29, 2007.

[24] L. Xie, H. Sundram, and M. Campbell. Event mining in multimeida streams. In *Proceedings of the IEEE*, 96(4):623–647, 2008.

[25] B. Yao and et al. Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool & benchmarks. In *EMMCVPR*, 2007.

[26] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *IEEE Int'l Conf. on CVPR*, vol. 2, pages 2057–2063, 2006.

[27] J. Yang, R. Yan and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188-197, 2007.

[28] X. Yu and et al. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In *ACM Multimedia*, pages 11–20, 2003.