Videography-Based Unconstrained Video Analysis

Kang Li, Sheng Li, *Student Member, IEEE*, Sangmin Oh, and Yun Fu, *Senior Member, IEEE*

***Abstract*— Video analysis and understanding play a central role in visual intelligence. In this paper, we aim to analyze unconstrained videos, by designing features and approaches to represent and analyze videography styles in the videos. Videography denotes the process of making videos. The unconstrained videos are defined as the long duration consumer videos that usually have diverse editing artifacts and significant complexity of contents. We propose to construct a *videography dictionary*, which can be utilized to represent every video clip as a sequence of videography words. In addition to semantic features, such as foreground object motion and camera motion, we also incorporate two novel interpretable features to characterize videography, including the scale information and the motion correlations. We then demonstrate that, by using statistical analysis methods, the unique videography signatures extracted from different events can be automatically identified. For real-world applications, we explore the use of videography analysis for three types of applications, including content-based video retrieval, video summarization (both visual and textual), and videography-based feature pooling. In the experiments, we evaluate the performance of our approach and other methods on a large-scale unconstrained video dataset, and show that the proposed approach significantly benefits video analysis in various ways.**

***Index Terms*— Videography analysis, video retrieval, video summarization, feature pooling.**

I. Introduction

AUTOMATIC understanding of visual content in unconstrained Internet video, such as those found on consumer video sharing sites (*e.g.*, YouTube and Metacafe), offers an interesting but very challenging task. These videos are particularly challenging because they contain very diverse content; they are captured under a variety of camera motion conditions (panning, zooming, translating); they are of highly variable length (from minutes to hours); and they are often heavily edited (*e.g.*, shot stitching and adding captions). Nowadays, a huge amount of unconstrained videos are captured by nonprofessional users, which makes the task of video understanding more challenging. As such, unconstrained videos are qualitatively very different and even more challenging than widely-used video datasets, such as the Hollywood dataset [18] or the YouTube Sports dataset [27], in which video clips contain fairly coherent single action occurring within a short duration. For example, some wedding videos from video sharing websites are more than an hour long and they are produced by stitching shots recorded separately across the entire wedding event. Each shot contains fairly different content, such as a panning camera capturing a party room filled with dancing guests, a series of stitched shots of each guest individually congratulating the wedding, or a shot that zooms in on the bride and groom. On the other hand, other wedding videos may be only minutes long, and only contain shots of the key events of the ceremony.

In this work, we present an approach for *unsupervised videography analysis* for this type of unconstrained video. Intuitively, each videography can be understood as a camera director's direction on a movie script, *e.g.*, "capture the running actress by panning the camera, to have her face appear at 20 percent size of the video". The idea is that different classes of video content will have different videography styles—the videography style of a wedding video should be different from a sports video—and so, the videography style should provide a valuable signal for automated content analysis. In this paper, we demonstrate the value of videography analysis for several important high-level tasks of intelligent video analysis. Specifically, we focus on three applications: 1) content-based video retrieval, 2) video summarization, and 3) videography-based feature pooling.

In our approach, we assume that there are diverse videography styles in unconstrained videos, which are discovered as a *videography dictionary* via unsupervised clustering on proposed features. Then, a video clip can be represented as a series of segments with varying videography words. For the underlying videography features, we extend conventional features such as camera motion and foreground (FG) object motion [10], [11], [17], [54] by incorporating two novel features: *motion correlation* and *scale information* (see Sec. III). To the best of our knowledge, our work is the first to address the explicit learning of a videography dictionary based on such a rich set of features beyond simple camera motions.

The overview of our proposed approaches is illustrated in Fig. 1. We first (step 1) decompose the long video clips into sequence of shots through shot boundary detection, then each shot is further divided into segments by motion-derived "camera operation boundaries". Meanwhile, the foreground and background motions can be separated. After chopping

Manuscript received July 8, 2016; revised January 17, 2017; accepted February 18, 2017. Date of publication March 5, 2017; date of current version March 27, 2017. This work was supported in part by the NSF IIS under Award 1651902, in part by NSF CNS under Award 1314484, in part by ONR under Award N00014-12-1-1028, in part by ONR Young Investigator under Award N00014-14-1-0484, and in part by the U.S. Army Research Office Young Investigator under Award W911NF-14-1-0218. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aydin Alatan.

K. Li and S. Li are with the Department of Electrical and Computer Engineering, College of Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: kangli@ccs.neu.edu; shengli@ece.neu.edu).

S. Oh was with Kitware, Inc., Clifton Park, NY USA. He is now with Faraday Future, Inc., Los Angeles, CA 90248 USA (e-mail: sangmin.oh1@gmail.com).

Y. Fu is with the Department of Electrical and Computer Engineering, College of Engineering and College of Computer and Information Science (Affiliated), Northeastern University, Boston, MA 02115 USA (e-mail: yunfu@ece.neu.edu).

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

Digital Object Identifier 10.1109/TIP.2017.2678800

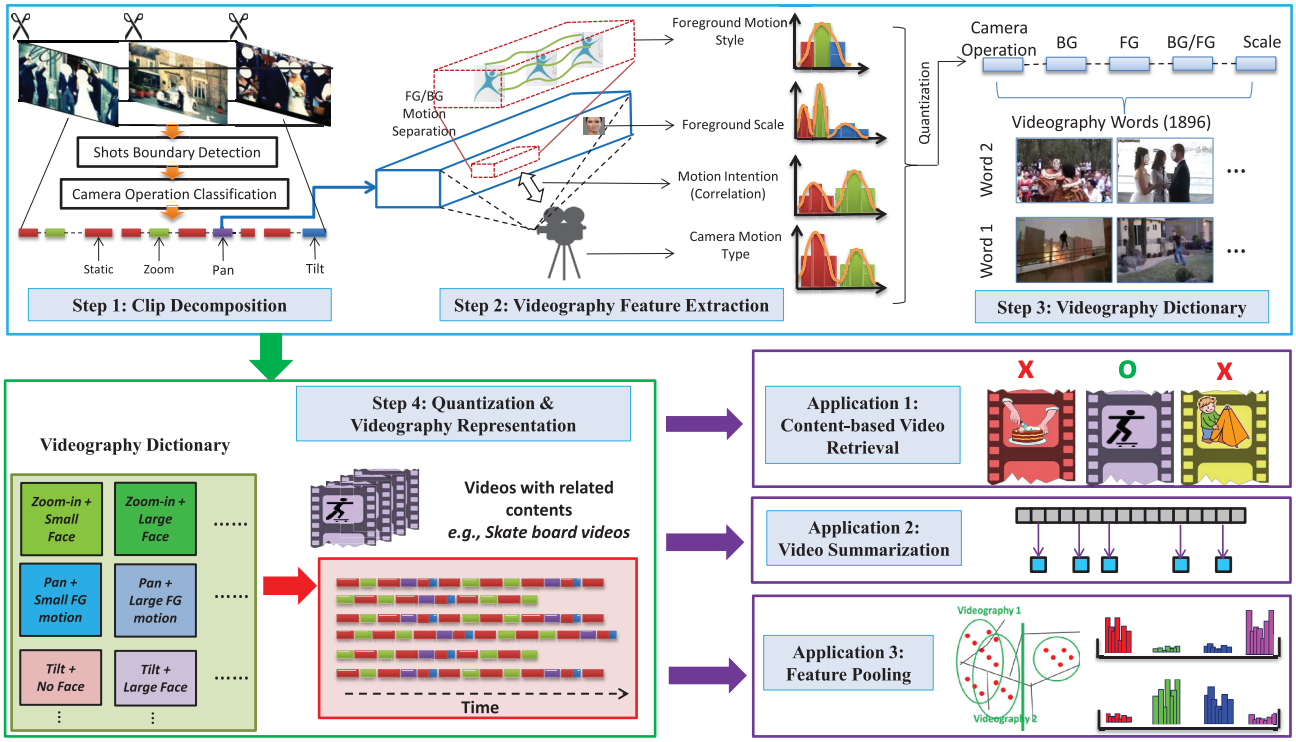

Fig. 1. Framework for videography analysis and applications for unconstrained videos. *Step 1*: Clip decomposition. *Step 2*: Videography feature extraction. *Step 3*: Videography dictionary construction from extracted features. *Step 4*: Test videos are quantized into videography word sequences and learning techniques are used to identify signature styles. *Step 5*: Applications. The learned models are used to provide (1) content-based video retrieval, (2) video summarization, and (3) feature pooling.

video into small pieces of relatively coherent content, we compute a series of features within each segment (as illustrated in Step 2 and described in Sec. III). Then we (step 3) cluster these features to develop a videography dictionary, and (step 4) quantize the segments into videography style words and learn the relationship between the style words and events. This is used for (application 1) content-based video retrieval, (application 2) content-adaptive video summarization, and (application 3) videography-based feature pooling.

For retrieval, we compare our approach with alternative methods on the NIST 2011 TRECVid Multimedia Event Detection video dataset (TRECVid MED 2011) [44] across 15 different diverse query collections, and show that the videography style does indeed add complementary information (Sec. V).

In addition, our adaptive summarization approach is different from the existing body of work relying on fixed rules (*e.g.*, [54]) in that our system optimizes summarization process to highlight the unique content of the given test videos (Sec. VI). As a mid-level semantic feature, videography can also be combined with other type of semantic features, such as Object Bank [23], to generate more detailed synopsis type of summary for the video (Sec. VI).

We also present a videography based feature pooling (VF-Pooling) method that utilizes the semantic information on segment-level, and further improves the performance of video recognition. As a novel semantic feature, videography has its unique perspective for video content analysis. Thus, it can be integrated with other features to generate more powerful representation. Specifically, by assigning a videography style label to each segment, our approach pre-categorizes segments into groups where each group corresponds to a videography style. In this way, we can build multiple descriptors for each clip, where each descriptor comes from a specific videography style group by averaging segment-level features that belong to that group. In a sense, these descriptors are *local* to the represented videography word. Our idea of VF-Pooling adopts the similar strategy as [15], inspired by recently introduced local pooling theory [4]. The general idea of local pooling is that pooling similar features separately in high-dimensional feature space would improve the overall representational power.

A. Contributions of this Work

This paper is a substantial extension of our previous work [22]. Compared to [22], we make the following extensions: (1) we add more technical details and discussions for videography features; (2) we add the textual-based video summarizations; (3) we add the feature pooling based on videography.

In this subsection, aspects of our proposed videography analysis framework are highlighted:

1) To the best of our knowledge, our work is the first to address the explicit learning of videography styles of unconstraint video. The idea is that different classes of video content will have different movie scripts which reflect the camera man or director's intention or direction. We believe that this type of information contains

TABLE I
ABBREVIATIONS

Abbreviation	Description
VW	videography word
MED	multimedia event detection
VD	videography dictionary
FG/BG	foreground/background
SBD	shot boundary detection
S/P/T/Z	static/pan/tilt/zoom
MI	mutual information
AP	average precision

interpretable semantics which can greatly help us understand the video content.

2) We demonstrate the value of videography analysis for several important high-level tasks of video analysis, including 1) content-based video retrieval, 2) video summarization (visual-based and textual-based), and 3) videography-based feature pooling.

To improve the readability of the paper, we have listed the abbreviations that used throughout this paper in Table I.

II. Related Work

The idea of representing videos as a series of segments based on motion and/or appearance characteristics has been explored to some extent, either as part of integrated systems [40], [49], [54] or on its own. Representations learning plays a central role in visual analytics [25], and effective video representations will greatly facilitate many video analysis tasks, such as human motion segmentation [24], and human activity recognition and prediction [20], [21], [26].

Some early works on video analysis, such as [40], utilizes videography heuristics with respect to keyframe selection. Our work shares a similar idea with them in terms of exploiting videography information. However, our approach explicitly learns videography styles, and can be applied to many high-level visual understanding tasks such as textual-based video summarization. Most systems, including this work, incorporate two main low-level processing steps: (a) shot boundary detection [35], [52], which is to find the boundaries between stitched shots, and (b) camera motion estimation within shots [36], [53], [54] to further decompose shots into finer sub-shot units based on evolving camera motion types.

In terms of videography modeling, the methods closest to our work are [49], [54]. In [54], a system capable of both summarization and retrieval was presented. The system is mostly based on hand-tuned distance metrics and rules to classify shots and videos into semantic categories, based on multiple features with heavy emphasis on appearance (*e.g.*, color and texture), and a few others such as simple camera motion primitives (i.e., Static/Pan/Tilt/Zoom; S/P/T/Z). In our retrieval experiments (Sec. V), we compare our new features with these simpler 4 types of camera motion primitives. It is worth noting that our work presents results primarily based on motion information without relying on appearance matching, and therefore provides a clearer understanding on the promise of motion-based videography modeling alone for high-level tasks. Additionally, since our approach is learning-based, the heavy burden of tuning system parameters is alleviated. In [49], the authors presented seven self-defined videography styles common in commercial movies, which are classified per shot based on features such as motion, appearance, and FG/BG separation; the videography quantization is based on supervised learning, and its use for summarization or retrieval is not studied. In contrast, our approach is unsupervised and does not require manually labeled training data for sub-shot classification, and hence can scale up for unconstrained videos with more complex videography styles beyond commercial movies.

Video summarization that has been well studied in multimedia community [6], [7], [13], [50] is formulated as a key frame extraction problem where change detection is commonly used based on appearance features such as color [49]. Different approaches which incorporate overall camera motion include [54]. However, both works adopted fixed rules for all videos. Above mentioned methods are visual summarizations which generate a "teaser" for the video. Recently, researchers start looking at the possibility to translate video content directly into human language [47]. The well-known challenge here is the longstanding semantic gap between low-level visual features and high-level semantic information.

The idea of soft feature pooling has been discussed in [4], however, it mainly considers low-level features for image classification. Our approach extends it from image analysis to video analysis at segment-level. Some recent works also explore video understanding at the segment-level, such as [31] and [5]. In [31], the distinctive temporal segments such as sub-actions are identified, in order to represent the complicated activities. This strategy works well for video data with fairly regularized structures, but may have limited performance on unconstrained consumer videos. In [5], the visual features extracted from video frames are clustered into different groups. During this process, a secondary feature (e.g., GIST [32]) is also involved to guide the clustering. However, [5] mainly utilizes the image-based features. Different from existing works, our approach is able to exploit diverse multimedia features (audio and visual), and can take advantages of temporal segments created based on some sort of semantic analysis.

Deep learning has attracted an increasing attention due to its impressive performance in various tasks [3]. In order to achieve better performance, the technical components in the proposed system, such as face detection and feature learning, could be replaced by the advanced deep learning methods [42]. But, in this paper, we mainly focus on the design of the system from a new perspective, and demonstrate the effectiveness of the videography feature.

III. Videography Features

For every input video, our approach applies two main processing steps to extract videography features, as illustrated in Fig. 1. First, a two-level motion analysis is conducted to decompose long clips into sequences of segments with coherent motion types (S/P/T/Z). Second, multiple features related to motion and scale patterns are measured from every segment, which are used to characterize videography. For both steps, we utilize densely computed KLT tracks [39] over the entire clips as the main basis for the derived features.

For the two-level decomposition, it is worth noting that we incorporate existing effective methods as part of our feature extraction module. In particular, we focus on: (a) developing novel techniques to enable high-level videography analysis; (b) its application for retrieval and summarization based on noisy videography quantization as intermediate representations. Shot boundary detection is believed to be largely solved; we adopt [52]. For background (BG) camera motion estimation, we extend [36], [53] to estimate three camera motion parameters (i.e., Pan/Tilt/Zoom; P/T/Z) from KLT tracks while simultaneously separating the tracks into FG/BG groups. We found that other approaches for FG/BG separation such as [11] are unsatisfactory for unconstrained videos, possibly due to the complex geometric scene structure in our data.

In the first phase, we use a shot boundary detection (SBD) algorithm which relies on the birth and death ratio of KLT tracks [52]. In detail, we developed two SBD modules, each one for two different styles of boundaries, namely: *Cut* (simple abrupt transition) and *Fade-Out-In* (common gradual transition), which account for majority of boundaries in videos. On labeled test data of 153 shot boundaries, the precision and recall are 0.95 and 0.98 for *Cut*, and 0.63 and 0.75 for *Fade-Out-In*, which are fairly good results.

Then, the second phase decomposes each shot further into sub-segments based on four camera motion types (S/P/T/Z). For unconstrained videos, camera motion estimation is challenging due to the complex interplay between the (apparent) motion of background (BG) and foreground (FG) objects, which need to be separated to yield accurate results. It is worthy to note that the background motion estimation problem mentioned in many existing papers [2] is quite different from ours. For example, [2] focuses on the BG/FG separation, not the BG/FG motion separation. In our scenario, we need to handle three problems: 1) FG/BG motion separation; 2) camera movement (four parameters) estimation; 3) FG motion estimation.

We adopt [36], [53] because of its proven performance on unconstrained videos and its advantage of solving FG/BG separation simultaneously. In detail, four standard types of camera motions are considered: pan (left or right), tilt (up or down), zoom (in or out) and static. We represent the image plane as a $K \times L$ regular grids, then fit the following affine camera model with four parameters at every frame:

$$\begin{bmatrix} V_{klx} \\ V_{kly} \end{bmatrix} = \begin{bmatrix} z_x E_{klx} \\ z_y E_{kly} \end{bmatrix} + \begin{bmatrix} p \\ t \end{bmatrix}. \quad (1)$$

Above, capitalized variables are known values where $V_{kl} = (V_{klx}, V_{kly})$ is the velocity vector of a block B_{kl}, and $E_{kl} = (E_{klx}, E_{kly})$ is the center of B_{kl}. Per-cell velocity V_{kl} is computed as the average from multiple tracks intersecting that cell. Lower-case variables are unknowns to be estimated, including zoom z, pan p, and tilt t. As suggested in [36], the block size is set to 8×8. Therefore, the dimensions of K and L vary with the size of video frame. In addition, we have tried other settings for K and L in a certain range, but the overall performance is not very sensitive to such settings.

We found that this grid-based formulation produces more reliable motion estimates by compensating frequently irregular spatial distribution of KLT tracks. The solution to Eq. 1 is straightforward by deriving a grid-version solution from [36], except that its accuracy will be guaranteed only when grid cells belonging to background are used as velocity observations. Accordingly, we solve it through iterative steps where FG tracks are identified as outliers under current camera motion estimates, and filtered out prior to updated camera motion estimation. It can be observed that our overall iterative camera motion estimation approach solves additional FG/BG separation problem simultaneously. Because per-frame solutions of Eq. 1 can be noisy, we use voting schemes across a time window to determine camera motions other than static. Specifically, we detect and classify the camera motion in a shot using experiential rules and thresholds. Let C_k be the camera parameter of the motion type k, and $C_k(s)$ be the motion parameter of C_k in frame s, $k \in \{pan, tilt, zoom\}$. A camera motion occurs if the following rules are satisfied:

- $|\sum_{s=i}^{j} C_k(s)| > T_k^{\text{sum}}$. The camera motion should occur noticeably, so the summation should exceed T_k^{sum}.
- $j - i + 1 > T_k^{\text{span}}$. The camera motion should occur continuously, so the duration should exceed T_k^{span}.
- $\frac{1}{j-i+1}|\sum_{s=i}^{j} C_k(s)| > T_k^{\text{avg}}$. The camera motion should occur uninterrupted and perceptibly, so the average of the summation should exceed T_k^{avg},

where i and j index the starting frame and ending frame of a candidate segment, respectively.

If the rules are satisfied, each type of camera motion is determined to occur in corresponding frames. For FG/BG KLT trajectory assignment, a similar voting method is applied along each trajectory by measuring the overlap portion of it with FG blocks.

As a result, KLT tracks are grouped into BG or FG, where BG group accounts for tracks mostly induced by camera motion and FG group as outliers from BG. Furthermore, to capture motion characteristics of FG objects accurately, FG tracks are motion-corrected by subtracting average BG motion. These are illustrated in Fig. 2(Top). Although FG/BG separation results are not perfect, the portion of mis-classified tracks is usually small, hence, unlikely to undermine the overall videography analysis.

Once segments are obtained, a set of videography features is extracted from every segment. In this work, we focus on visual features related to *motion* and *scale*: (1) camera motion type (S/P/T/Z), (2) FG and (3) BG motion, (4) correlations between FG/BG motion, and (5) the scale of foreground. The videography feature is finally represented as a 10 dimensional feature vector. [1]

For FG and BG motion, the average motion within a segment is normalized w.r.t. the video width, to cope with video

[1]The detailed information of each dimension in a videography feature vector: (1) PAN: 1/pan-right, -1/pan-left, 0/no-pan; (2) TILT:1/tilt-up, -1/tilt-down, 0/no-tilt; (3) ZOOM: 1/zoom-in, -1/zoom-out, 0/no-zoom; (4) static: 1/yes, 0/no; (5) Background motion magnitude (normalized to percentage of the video width); (6) Foreground motion magnitude (normalized to percentage of the video width); (7) Face scale (normalized to percentage of the video width); (8) Face count; (9) FG/BG correlation (percentage of FG tracks that have "same" direction with camera); (10) FG tracks ratio (percentage of tracks that belong to foreground).

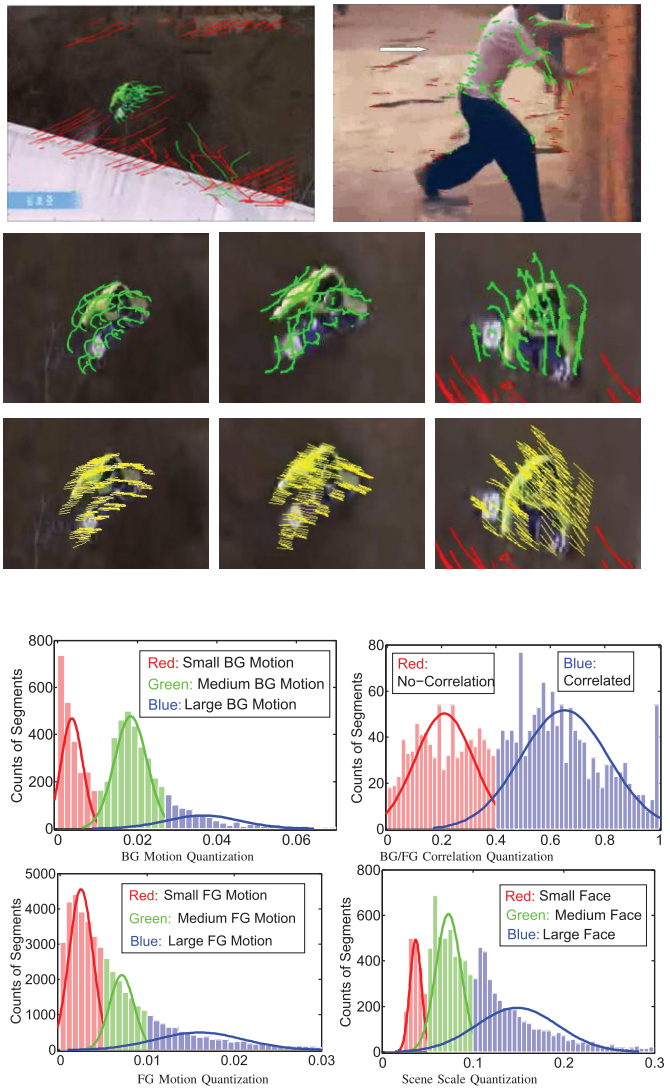

Fig. 2. Videography feature extraction. (Top) Camera motion estimation with FG/BG separation. (Middle) Original FG motion (green) is corrected (yellow). (Bottom) Distribution of extracted videography features, and a clustering-based quantization.

clips with varying sizes. Our novel FG/BG correlation feature is motivated by the fact that similar camera motion may be invoked by different intentions, *e.g.*, tracking or simply switch of focus. The magnitudes of FG/BG correlation are measured by the normalized sum of inner product between FG tracks and average BG motion. We also include scales of FG objects as another distinctive feature for videography. For example, clips with close-up shots of faces are very different from clips which contain far-away shots of pedestrians. Because the estimation of scale is a very challenging problem, in this work, we used the bounding box sizes of face detections produced by off-the-shelf systems (*e.g.*, [48]) as a proxy for scale estimates. In detail, average face size within a segment (normalized by the video height) is used to represent the scale. For example, face scale of 0.2 indicates that the average size of faces occupies about 20 percent of the image height. We also notice that a lot of advanced face detection methods have been proposed in recent years, such as [55]. However, as we are dealing with a huge unconstrained data set with low-quality videos, the classical and efficient face detectors like the Viola-Jones face detector [48] can already obtain comparable results than the most recent methods in the unconstrained scenario. Moreover, our paper mainly focuses on presenting a novel framework to extract videography features for a series of high-level video analysis tasks, and it's quite flexible to replace any building blocks with other appropriate methods (e.g., human body detection [37]) in practice.

IV. Videography Dictionary and Analysis

Once videography features are obtained from segments, they are grouped to form videography dictionary (VD) shown in Fig. 1. The computed VD will be used to quantize video clips into sets of videography words (VWs), as shown in Fig. 1.

For our experiments, we extracted the above-mentioned videography features from a training video dataset, which consists of roughly 2000 unconstrained videos (∼80 hours total), where 29 segments are found per clip on average. The overall distribution of the extracted features is shown in Fig. 2(Bottom), where the multi-modal characteristics in most videography features (except FG motion) can be observed. Such patterns indicate that there are indeed regularized videography patterns in videos.

We have explored two different methods for developing the dictionary: (1) concatenated and (2) joint learning. In the first *concatenated* learning, each feature dimension is quantized individually, and then all the features are concatenated to form VD in a combinatoric manner. Straightforwardly, the first feature dimension of camera motion type has four quantization values of S/P/T/Z. We quantize the remaining features individually, based on an empirical analysis of the data on the training set. In particular, we used regular intervals when quantizing the features. As illustrated in Fig. 2(Bottom), the BG/FG motion is separately quantized into *small/medium/large*; the FG/BG correlation is quantized into *correlation* or *no-correlation*; and the scale is quantized into *no-face/small/medium/large*. In particular, the face scale feature is quantified to 0 if no faces are detected from video. The video words are then formed by concatenating these values. This procedure creates $4 \times 3 \times 3 \times 2 \times 4 = 288$ possible video words.

Our analysis of the distribution of the resulting VD shows that, interestingly, only ∼40% of the words are actually observed in the data, indicating that only a subset of combinations of feature quantizations are present, *e.g.*, a combination such as zoom-in, large FG and BG motion, no correlation, and large scale actually does not appear. Furthermore, if we eliminate rare words which have fewer than ten occurrences, we are left with only 82 unique videography words, over a dataset of 80 hours of unconstrained video. Such observation provides an insight that there are fairly regularized patterns in how people capture videos, regardless of content. To the best of our knowledge, this is the first study that provides automated analysis on characteristics of videography styles on unconstrained Internet videos.

In the second method of *joint learning* for developing the dictionary, we again quantize the motion type into the same

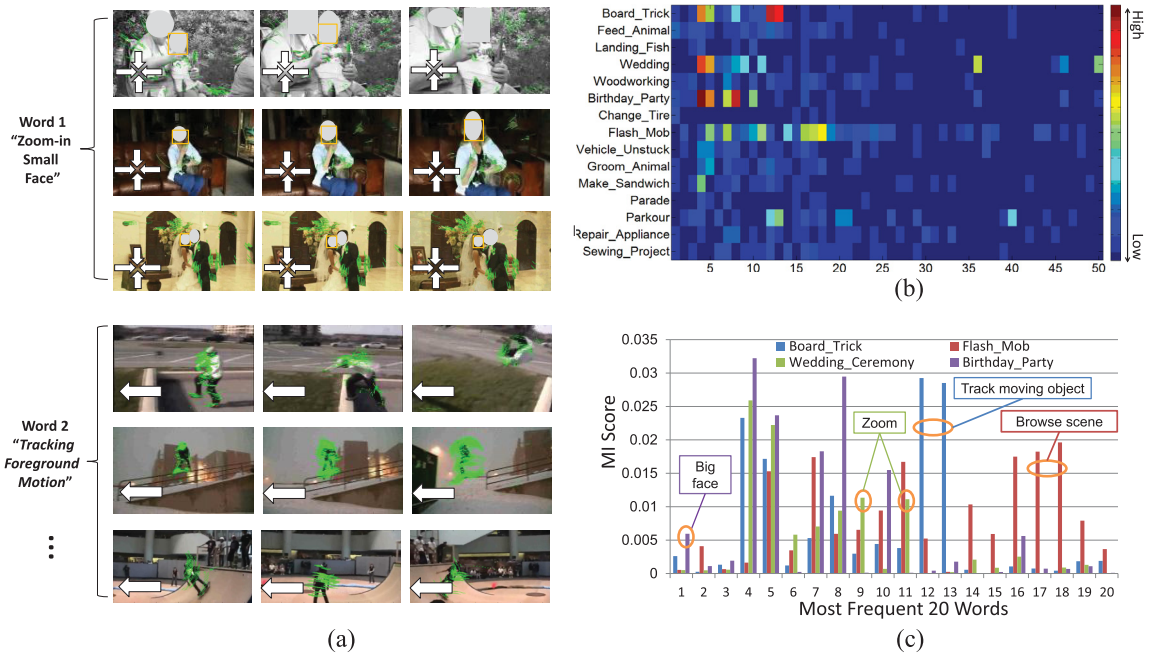

Fig. 3. (a) Videography word examples. (b) Mutual information between different event classes and most frequent 50 VWs. (c) Qualitative analysis on 4 event classes.

four values (S/P/T/Z). However, for each motion type, we perform K-means clustering on the remaining four-dimension continuous vector space formed by concatenating the four raw feature types (FG motion, BG motion, amount of correlation, size of face). In our experiments, we chose K=30, which yields $4 \times 30 = 120$ video words. We used a smaller number of clusters because of the observation that many of the video words from the first method were actually not used.

In addition, the K-means method here can be replaced by some advanced dictionary learning methods, such as K-SVD [1], in order to learn informative bases for representation.

Once VDs are obtained, we can examine their accuracy as a macro feature type by examining the sample video segments in each word cluster. Example segments belonging to two sample videography word clusters are shown in Fig. 3(a), along with the detected visual features overlaid on images to show more details, including camera motion (left bottom arrows), compensated FG motion (green tracks), and face detections (orange boxes).[2] The textual descriptions of both words were produced manually, by looking at both the feature vector values and the grouped segments. It can be observed that segments with highly related content are successfully grouped into the same VWs. In particular, it is worth noting that in the second example, similar segments are grouped together correctly, even though faces are not detected due to the challenging imaging conditions. We have manually examined 10 VWs by drawing 30 segment samples each and concluded that, on average, 88% of segments from the same VWs show perceptually identical videography.

We qualitatively compare the videography words decided by the joint learning strategy and the concatenated learning strategy. By using the joint learning, the first three rows are characterized as videography words "Zoom-in Small Face", and the bottom three rows as "Tracking Foreground Motion", as shown in Fig. 3(a). In the case of concatenated learning, the decided words for the first three rows are "Zoom + Small (BG motion) + Small (FG motion) + Correlation + Small Face". And the words for the last three rows are "Tilt + Small (BG motion) + Large (FG motion) + No Correlation + No-face/Small Face". Clearly, the joint learning is an unsupervised data-driven strategy, and could truthfully represent the videography style of the grouped video segments.

We also conducted analysis on the correlations between VWs and particular visual content, so called *events*. By events, we mean semantic content classes captured in videos, such as *Flash mob* or *Birthday party* (defined further in [44]). This notion of analyzing or learning about videography of videos containing the same events is illustrated in Fig. 3(b,c). Specifically, we measured the mutual information (MI) between each word and each event. A high MI score indicates that a word is discriminative for the corresponding event. Our results are summarized in Fig. 3(b) where MI between every event and top 50 most frequent VWs are shown. It can be observed that, for a particular event, there are certain *signature* VWs. More detailed analysis is shown for four event types and top 20 words, in Fig. 3(c). In particular, this analysis provides insight on how different events are captured with different styles. For example, it shows that event *Board trick* has a strong style of *tracking moving object*; event *Flash mob* has a strong style of *browsing scenes*; event *Wedding ceremony* shows frequent *zooming*; and event *Birthday party* shows frequent *facial close-up*. This observation on discriminative correlations suggests that videography analysis can actually be used for challenging tasks such as retrieval (Sec. V) and summarization (Sec. VI).

[2]In this work, faces are intentionally occluded in this figure for privacy.

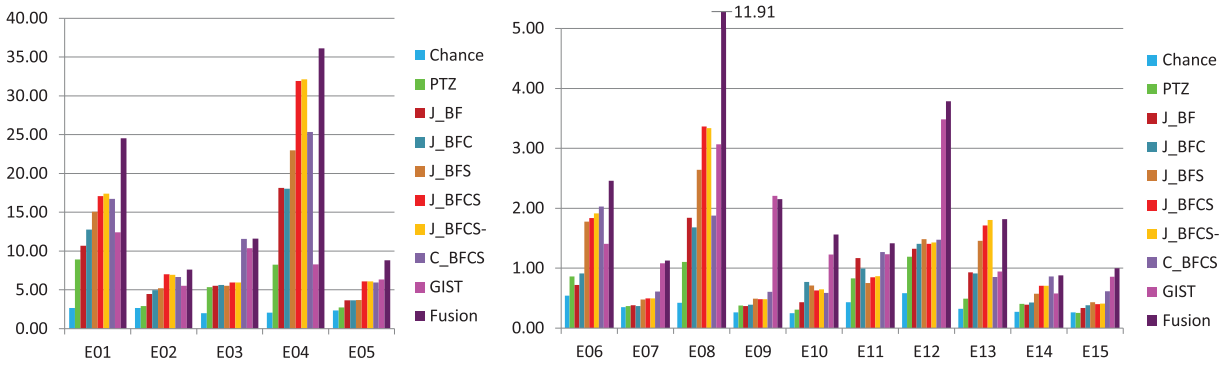

Fig. 4. Average Precision (%) of video retrieval results on MED corpus, for 15 events: ***(E01) Board trick***, *(E02) Feeding animal, (E03) Fishing,* ***(E04) Wedding***, *(E05) Working wood project,* ***(E06) Birthday party***, *(E07) Change vehicle tire,* ***(E08) Flash mob***, *(E09) Getting vehicle unstuck, (E10) Groom animal, (E11) Make sandwich,* ***(E12) Parade***, ***(E13) Parkour***, *(E14), Repair appliance, and (E15) Sewing project.*

TABLE II

MEAN AVERAGE PRECISION (%) OF VIDEO RETRIEVAL RESULTS ON MED CORPUS, FOR TWO SEPARATE TEST DATASETS OF EVENTS 1-5 AND EVENTS 6-15 RESPECTIVELY. FUSION RESULTS ARE OBTAINED BY COMBINING J_BFCS AND GIST. THE RESULTS WITH DYNAMIC EVENTS ONLY ARE MARKED WITH (D), WHICH INCLUDE EVENTS: E01, E04, E06, E08, E12, AND E13. **BOLD** FONTS DENOTE THE BEST RESULT, AND *Italic* FONTS DENOTE THE SECOND-BEST RESULT

	Baselines			Ours						
mAP	Chance	LSTM [51]	Corr-LDA [12]	PTZ	J_BFCS	J_BFCS(D)	GIST	GIST(D)	Fusion	Fusion(D)
E01-E05	2.34	4.2	24.3	5.63	13.61	24.50	8.57	10.34	*17.74*	**30.35**
E06-E15	0.37	4.75	N/A	0.62	1.19	2.08	1.61	2.22	*2.81*	**4.99**

V. APPLICATION FOR VIDEO RETRIEVAL

In this section, we present our approach and experimental results for videography-based video retrieval. In detail, we computed videography word bag-of-word (VW-BoW) representations, where per-clip unigram features are built from set of VWs (regardless of temporal ordering), for every clip. The goals are to examine (1) how well the proposed VW-BoW feature can perform in retrieval tasks by itself, compared to other alternatives and with detailed studies on the contribution of each videography feature component, and (2) whether our approach offers a useful modality to capture characteristics of video belonging to high-level event classes, in comparison to other macro-level features such as GIST [32].

For dataset, we use TRECVid 2011 multimedia event detection (MED) corpus [44] as our data, due to its large size, realistic content variability, and existing clip-level annotations for 15 different event classes. Both the scale and complexity of the dataset are beyond the widely-used datasets [18], [27]. Clips are frequently captured in unconstrained lighting and camera motion conditions, exhibiting diverse degrees of encoding artifacts and severe background clutter, and heavily edited by owners using shot stitching, caption embedding, etc. For training data, we use "Part-1 training data" (called event kits), which consists of videos from 15 different event classes of 137 clips per class on average (total 2061 clips) with average duration of 4.2 minutes. From these training data, our VDs are computed by selecting the best run out of 100 K-means clustering, and later used for test data. The 15 event types are enlisted in the caption of Fig. 4, with events frequently exhibiting complex camera motion marked in bold faces. For test data, MED corpus provides two different subsets, "Part-1 DEV-T" for the first 5 event classes, and "MED11TEST" for the remaining 10 event classes, with 4292 and 32061 total clips respectively. Both test datasets contain large amount of negative clips which do not belong to any of the target event classes, consequently, they serve as realistic test-bed for retrieval experiments. The positive examples in the two test datasets only constitute 2.34% and 0.37% on average per class respectively.

Our retrieval experiments are conducted using one-vs-all SVM classifiers, parameters of which are tuned via cross-validation. The overall results are summarized in Fig. 4 and Table II, where several experiments are conducted. As performance metrics, average precision (AP) is used. It is worth noting that APs for E06-E15 are lower than E01-E05, because the relative ratio of negative samples in the test dataset for E06-E15 is about 10 times higher. In detail, *Chance* denotes random retrieval and *PTZ* denotes the use of four-dimensional BoWs of discrete camera motion types only (*e.g.*, S/P/T/Z) without detailed videography features, as comparative methods [54]. The variations of our approaches are marked using abbreviations where J and C denote joint or concatenated VD learning, described in Sec. IV. Additionally, B, F, C, S indicate the inclusion of BG motion, FG motion, BG/FG correlation, and scale respectively, during VD learning. These experiments have been conducted to examine the usefulness of each videography feature for retrieval. The minus sign '$-$' indicates that the VD has been pruned by filtering out VWs with low MI scores per event type. For all the experiments with BoW-type features, histogram intersection kernel (HIK) was used for SVM training and testing. In addition, we

TABLE III

MEAN AVERAGE PRECISION (%) OF VIDEO RETRIEVAL RESULTS ON MED CORPUS. THE TABLE BELOW ENLISTS ALL THE DETAILED QUANTITATIVE RESULTS INCLUDED IN THE PAPER. FOR DETAILED DESCRIPTION AND EXPERIMENT ACRONYMS, PLEASE REFER TO THE MANUSCRIPT

	E01	E02	E03	E04	E05	mAP(D)	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15	mAP(D)
Chance	2.66	2.66	2.00	2.07	2.33	2.37	0.54	0.35	0.42	0.26	0.25	0.43	0.58	0.32	0.27	0.26	0.47
S/P/T/Z	8.92	2.90	5.35	8.26	2.74	8.59	0.86	0.37	1.10	0.37	0.31	0.83	1.19	0.49	0.40	0.25	0.91
J-BGC	12.77	4.96	5.64	18.02	3.65	15.10	0.91	0.37	1.68	0.39	0.77	1.00	1.41	0.91	0.42	0.38	1.23
J-BGS	15.08	5.21	5.53	22.99	3.68	19.03	1.78	0.48	2.64	0.49	0.71	0.75	1.48	1.46	0.57	0.43	1.84
J-BGCS	2.66	2.66	2.00	2.07	2.33	2.37	0.54	0.35	0.42	0.26	0.25	0.43	0.58	0.32	0.27	0.26	0.47
J-BGCS-	8.92	2.90	5.35	8.26	2.74	8.59	0.86	0.37	1.10	0.37	0.31	0.83	1.19	0.49	0.40	0.25	0.91
C-BFCS	12.77	4.96	5.64	18.02	3.65	15.10	0.91	0.37	1.68	0.39	0.77	1.00	1.41	0.91	0.42	0.38	1.23
GIST	15.08	5.21	5.53	22.99	3.68	19.03	1.78	0.48	2.64	0.49	0.71	0.75	1.48	1.46	0.57	0.43	1.84
Fusion	24.55	7.59	11.62	36.14	8.80	30.35	2.46	1.33	11.91	2.15	1.56	1.41	3.78	1.82	0.88	1.00	4.99

compare our approach with two recently proposed video event detection and video retrieval methods that are also conducted on the TRECVid 2011 MED dataset, including long short term memory networks (LSTM) [51] and correspondence-latent Dirichlet allocation (corr-LDA) [12]. Table II shows that the proposed approach with fusion strategy achieves higher mAP than LSTM and corr-LDA.

[32] shows the results using GIST features with linear SVMs. Because GIST is a per-image feature, GIST features are computed on frames extracted from labeled video clips. Then, one-vs-all SVMs were trained on image features using clip labels. For testing, SVMs are applied on extracted images, then, scores were averaged to produce a clip-level score. Apparently, VWs and GIST capture very distinct signals from data. Accordingly, in the experiment marked as *Fusion*, we have further explored whether fusion of two modalities can lead to further improvement, which will show whether these two feature types are complementary. For fusion, we have used the approach of "late fusion" (*e.g.*, [14]) where we have used the weighted sum of two classifiers as the fusion score. Among VW-based approaches, J_BFCS was used because it has been shown to provide best performance, and weights were determined by cross validation where equal weights of $<0.5, 0.5>$ were found to be best.

Overall, we can observe that VWs clearly provide advantage over the conventional simpler alternative of using camera types only, *i.e.*, *PTZ*. From Fig. 4, it can also be observed that every videography feature contributes towards improving performance. Between joint and concatenated VD learning, joint learning shows superior performance overall, possibly due to the data-driven construction of the dictionary which avoids many empty (or rare) VWs in concatenated learning. However, pruning VWs by MI scores does not seem to necessarily boost performance. Table III shows mean average precision (mAP) for key experiments in Fig. 4 on two test datasets. It can be observed that motion-based macro feature such as videography can outperform GIST for E01-E05 in "Part-1 DEV-T" set, and E06, E08, E11, E13, E14 in "MED11TEST" set. More importantly, the fusion results are much better than either approach, indicating that two feature types are complimentary. Table III also shows mAPs for dynamic events only, where we observe big boost in performance for VWs. Dynamic events mean that the there are usually significant changes between different shots in the videos. Interestingly, the event classes which show clear discriminative correlation with VWs in Fig. 3(b) are dynamic events, and they also show more advantages when VWs are used for retrieval. We also notice that some recent methods that utilize Fisher vectors [33], [41] of spatio-temporal visual words achieved impressive results on the video retrieval task, which demonstrates the effectiveness of advanced features like Fisher vectors. The videography features proposed could be fused with other features in real-world applications.

VI. APPLICATION FOR VIDEO SUMMARIZATION

With the huge amount of video content data available, it's essential to find important or interested contents efficiently, but unfortunately, it's impossible to edit those videos for a concise version manually. Thus, automatic video summarization techniques are badly needed, and construct the basis for many important video analysis tasks, such as those for the intelligence and security purpose. In this section, we present our videography-aware adaptive summarization methods.

Conceptually, content summarization is a kind of information abstraction either by selecting the most informative portions of content, or by refining the content into natural language description. In terms of video summarization, we call the former "visual summarization". For example, a movie teaser would be an example of this type of summary. The latter one is actually a "recounting" process to generate a series of textual notes which best represents the event happened in the video. This type of summary is a cross-modal information abstraction, and consequently more difficult and always less accurate.

In this paper, we demonstrate how to utilize videography analysis as an effective way to generate both visual and textual summarization. For visual summarization, we stitch key frames highlighted by event-relevant videography styles. For textual summarization, we combine the mid-level semantics of videography with another mid-level semantic feature, Object Bank, to generate key phrases level of summarization.

A. Visual Summarization

Visual summarization is designed to highlight the segments with distinctive videography styles for particular events. Our novel insight is that identification of segments from videos where cameramen are systematically exhibiting distinctive videography styles for particular events will provide unique

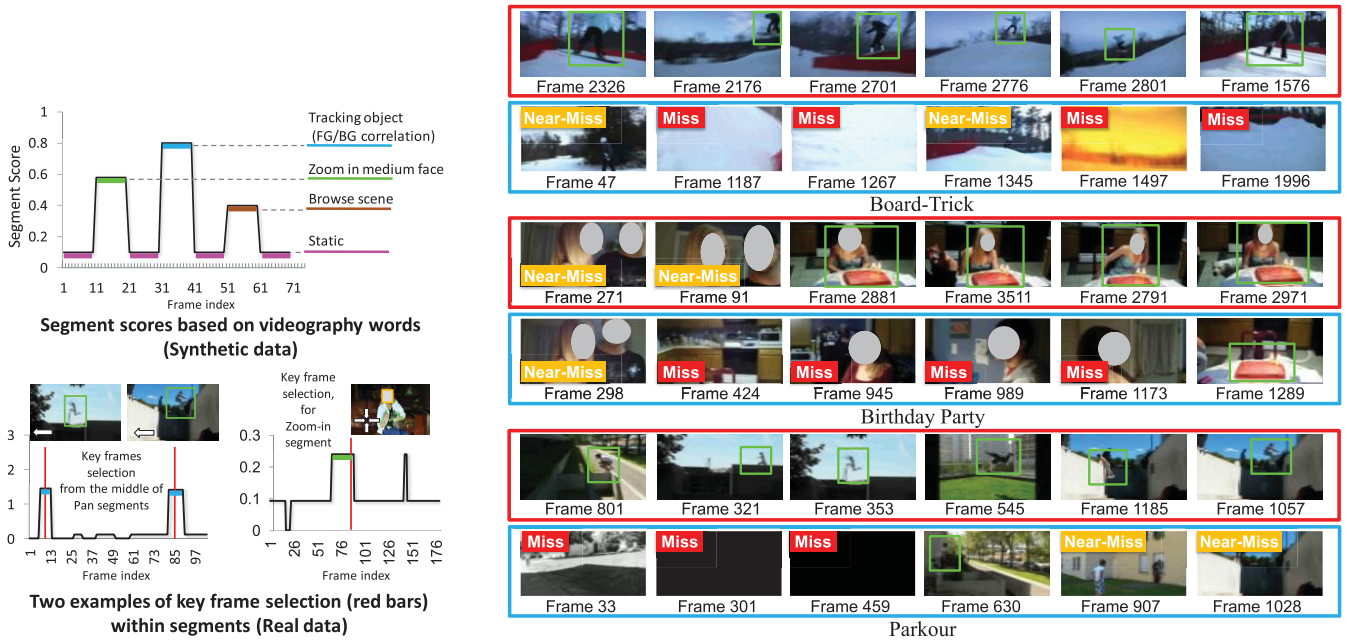

Fig. 5. Videography-aware adaptive summarization. (Left) Segment scores are based on MIs of corresponding VWs. Frames are selected at designated relative location within segments. (Right) Three summarization results by this work (red rows) and baseline (blue rows). Detected FG regions (green) and human judgements on relevance of key frames (good:none, near-miss: yellow, miss: red) to associated events are marked on each image.

summarization, assuming that such segments are strongly correlated with the major region of interest. While many works deliberately avoid the use of segments with motion due to complexity, *e.g.*, [30], such segments can be indeed crucial to characterize dynamic contents in videos exhibiting frequent camera motion, frequently recorded by mobile devices.

In our approach, frames are extracted by two step procedures, as illustrated in Fig. 5 (Left). First, key segments are selected based on segment scores, with optional weighted sampling scheme in case there are more number of segments than the desired number of key frames. For segment scores, MI scores have been used.[3] Then, key frames are extracted, one per selected segment. In particular, our novel innovation is that frames are designed to be extracted from different relative location within each segment based on their videography. Two different types of key frame selection mechanisms were used: frames are selected (1) in the middle of segments when videography is either stationary or indicates FG/BG correlation (to capture peak of FG motion), and (2) at either end of segments when the videography indicates P/T/Z without FG/BG correlation (to capture the destination of the shifting attention).

Qualitative summarization results are shown in Fig. 5 (Right), where frames extracted from same videos by our proposed method (red rows) and a conventional baseline (blue rows) are compared, for three different event classes. The results of the baseline method were obtained by extracting frames with highest scores based on color histogram changes, which is very common. It can be observed that our method is very effective in identifying unique contents from clips.

[3]Without event labels, term frequency inverse document frequency (tf-idf) scores [28] can be used instead.

TABLE IV
EVALUATION OF VISUAL SUMMARIZATION

Method	Meaningful Summary Duration (MSD)
SD+SVM	16.7
KVS	12.5
Ours	12.9

In particular, most extracted frames contain important visual moments when the FG people are at the peak of their action or camera focus, such as skilled jumps or before blowing a birthday cake candle. On the other hand, results by the baseline tend to include frames that just exhibit strong changing background or even black frames around the captions inserted by users. Overall, we observe that the proposed method can generate good visual summaries, especially for clips which contain complex camera motions.

We also perform quantitative evaluations for video summarization. By following the evaluation protocols and evaluation metric in [34], we compare our approach with the baselines, including shot detection (SD) with SVM, and kernel video summarization (KVS) [34]. The evaluation metric is meaningful summary duration (MSD), and a good summarization is corresponding a low MSD score. Table IV shows that our approach could achieve comparable performance with KVS.

B. Textual Summarization

Natural language as the most sophisticated information vehicle has been the core of AI research for decades. There exist many challenges in developing automatic video contents translation systems [8], [9], [19]. One well-known challenge is the long-standing semantic gap between computable low-level features and semantic information that they encode.

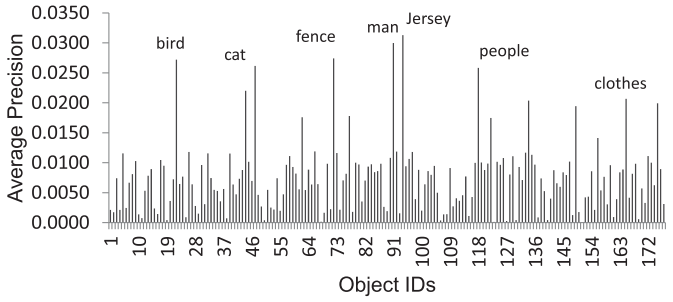

Fig. 6. The average precision values of all objects with respect to event groom animal.

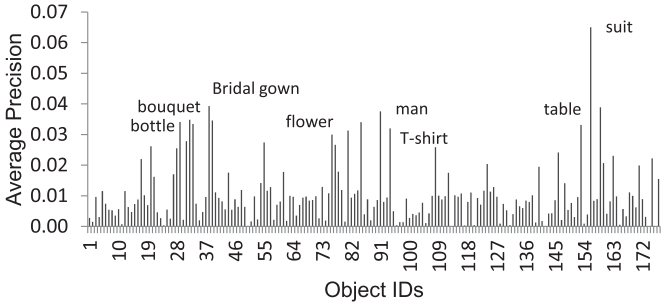

Fig. 7. The average precision values of all objects with respect to event wedding.

For unconstrained consumer video, though sentence-level textual summarization is still an extremely challenging task at current level of AI technology, the word-level or phrase-level of "video recounting" have become more and more promising due to a series of semantic features recently introduced in the literature. The proposed videography feature is also a mid-level feature with a clear semantic meaning. The interpretability of videography makes it a good complement to other semantic features, such as Object Bank. Specifically, in our approach, we utilize the detected object labels and videography style labels to generate a phrase-level summary. In particular, the template based language models are employed to generate phrases [43]. For example, we can generate a phrase "the camera zoom into a bride's face" by combining videography label "zoom in a middle-size face" with the detected object "bride". Or, we can generate a phrase "the camera is panning left tracking a moving person" by combining videography label "pan to the left tracking an object" (strong FG/BG correlation) with the detected object "human". Because videography is a motion-driven feature, it can add dynamic information about the video content, which provides a lively graphic description of the scene.

Object Bank feature in total has 177 object detectors, where each object has 6 different scales and 2 views. We still conduct experiments on the TRECVid MED 2011 dataset. First, we evaluate the discriminability of each object with respect to each event by ranking object labels according to their average precision (AP) scores for retrieval task. Figs. 6 and 7 show the detailed average precision values for all object classes with respect to event E01 to E05 in the TRECVid MED 2011 corpus, respectively. Fig. 8 shows the results of Object Bank [23], the Sequence to Sequence based Video to Text (S2VT) method [46], and the phrase-level textual summarization results of our method. S2VT is the state-of-the-art method that automatically generating descriptions from videos. Fig. 8 shows that our method and S2VT provides complementary descriptions of the video, as they are motivated from different perspectives. In particular, S2VT generates accurate descriptions by virtue of deep feature learning, while our method explicitly shows the transition of video by exploiting the camera motion information.

Fig. 8. Textual summarization of the video samples.

VII. Videography-Based Feature Pooling

Though many video retrieval systems (e.g., [14], [29]) have successfully adopted clip-level representations through global pooling strategy, the detailed temporal structure, especially for long-duration unconstrained consumer videos, has been ignored completely. Recent efforts [5] start looking at this problem and exploiting the potential of segment level descriptors. In this section, we will present a videography based feature pooling (VF-Pooling) approach that leverages the segment-level semantics and improves the model performance. As a novel semantic feature, videography has its unique perspective for video content analysis. Thus, how to effectively integrate multiple semantic features together becomes quite intriguing.

Specifically, by assigning a videography style label to each segment, our approach pre-categorizes segments into groups where each group corresponds to a videography style. In this way, we can build multiple descriptors for each clip, where each descriptor comes from a specific videography style group

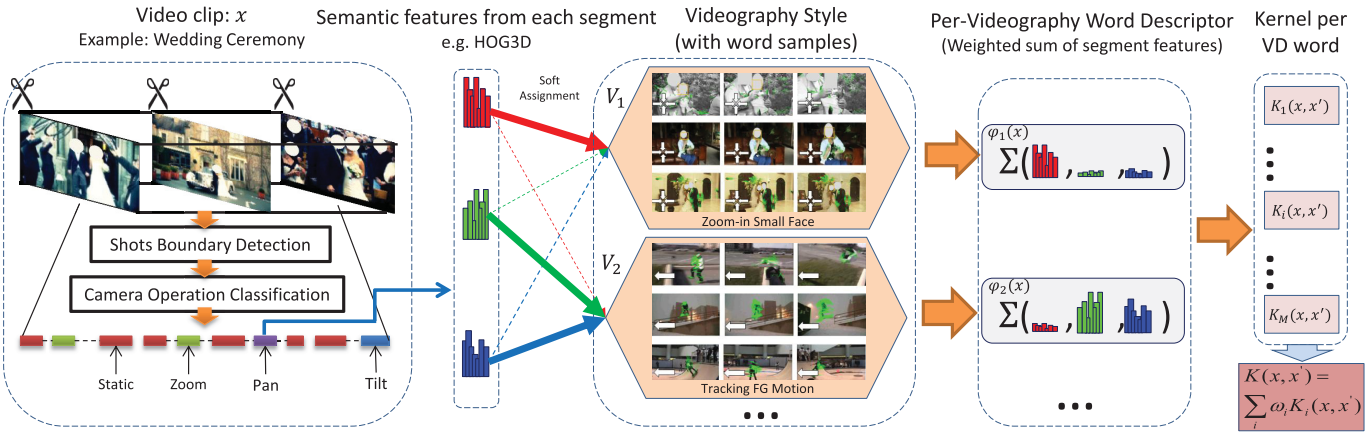

Fig. 9. Videography based feature pooling.

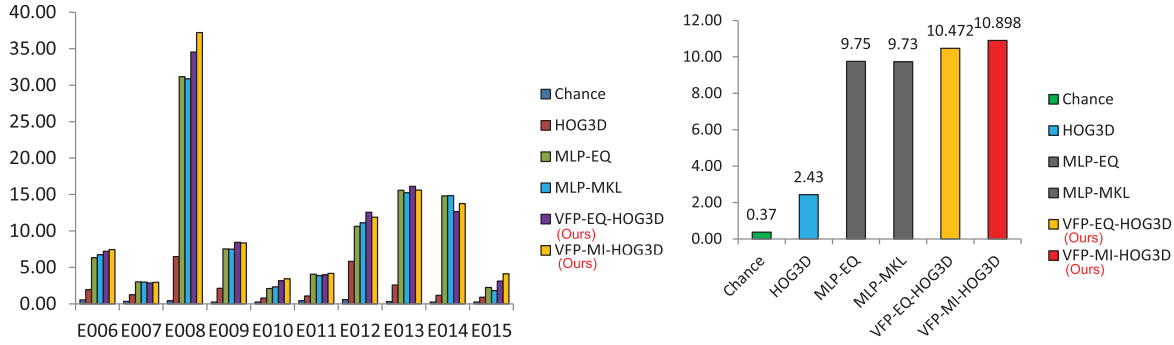

Fig. 10. Videography based feature pooling results. The metric AP(%) is used to evaluate baseline methods and the proposed VF-Pooling frameworks.

by averaging segment-level features belong to that group. In a sense, these descriptors are *local* to the represented videography word. Our idea of VF-Pooling adopts the similar strategy as [15], inspired by recently introduced local pooling theory [4]. The general idea of local pooling is that pooling similar features separately in high-dimensional feature space would improve the overall representational power.

The VF-Pooling framework is illustrated in Fig. 9. First, according to our videography analysis, we decompose each video clip into a set of segments, and then soft-assign every segment some videography style labels (VD word). A large assignment value indicates that strong videography style happened in current segment. Soft-assignment is important because when we build videography dictionary, we applied unsupervised clustering method, e.g. K-means, which may lead to arbitrary space partitioning. Then, each segment will be represented by a given feature. In our evaluation, we use feature HOG3D [16] as descriptor for each segment. The extracted feature descriptors for videos could be regarded as projections from a video clip to the videography space. After building multiple feature descriptors, kernelization can be separately applied to measure the similarity between different video clips w.r.t. each videography style. Finally, multiple kernel fusion techniques are applied to provide improved discriminant power for video event classification.

In detail, let $x_a = \{x_a^i | \, x_a^i \in R^{D1}, \, 1 \leq i \leq n\}$ be a training sample, where x^i is a $D1$-dimensional videography feature representation for the i-th segment, and n denotes the total number of segments in a video clip a. Based on videography features, we learn a VD dictionary by K-means clustering. Then, each centroid of the K-means clusters represents a videography style in the archive. Assuming the VD dictionary size is M, each VD word is represented by a centroid v_j, then we can represent the videography style set as $V = \{v_j | \, v_j \in R^{D1}, \, 1 \leq j \leq M\}$. Let $y_a = \{y_a^i | \, y_a^i \in R^{D2}, \, 1 \leq i \leq n\}$ be the $D2$-dimensional HOG3D feature representation for the i-th segment. Then feature descriptor $\varphi_j(y_a)$ of a video a corresponding to the j-th videography style is formulated as a weighted representation calculated for the entire video segments $\{x^1, x^2, \cdots, x^n\}$, with corresponding soft-assignment weights as

$$\varphi_j(y_a) = \frac{1}{n} \sum_{i=1}^{n} \omega_j(S_j, \, y_a^i) \cdot y_a^i, \tag{2}$$

where $\omega_j(S_j, \, y_a^i)$ denotes a soft-weight assignment function between the j-th VD word S_j and the feature representation of video segment y_a^i.

For evaluation, we select the histogram intersection kernel (HIK) SVMs to train classifiers. The kernel between a pair of video samples is calculated by integrating all the kernels calculated for each videography style. In addition, our work explores several different variations of kernel fusion, such as average weighted and Mutual Information weighted. Based on our previous discovery, we notice that certain videography styles are more discriminative for a particular event. Exploiting mutual information makes it possible to assign

TABLE V

MEAN AVERAGE PRECISION (%) OF VIDEO RETRIEVAL RESULTS ON MED CORPUS BY USING VIDEOGRAPHY BASED FEATURE POOLING TECHNIQUES. HERE WE USED HOG3D FEATURES FOR EVALUATION. THE TABLE BELOW ENLISTS ALL THE DETAILED QUANTITATIVE RESULTS INCLUDED IN THE PAPER. FOR DETAILED DESCRIPTION AND EXPERIMENT ACRONYMS, PLEASE REFER TO THE MANUSCRIPT. **BOLD** FONTS DENOTE THE BEST RESULT FOR EACH EVENT

EventID	E006	E007	E008	E009	E010	E011	E012	E013	E014	E015	mAP
Chance	0.54	0.35	0.42	0.26	0.25	0.43	0.58	0.32	0.27	0.26	0.37
HOG3D	1.97	1.25	6.48	2.15	0.81	1.1	5.83	2.58	1.18	0.92	2.43
MLP-EQ	6.34	3.01	31.16	7.54	2.11	4.07	10.63	15.57	14.81	2.25	9.75
MLP-MKL	6.74	2.98	30.87	7.50	2.34	3.86	11.13	15.25	14.84	1.82	9.73
VFP-EQ-HOG3D (Ours)	7.23	2.88	34.54	**8.45**	3.15	3.98	**12.58**	**16.12**	12.65	3.14	10.47
VFP-MI-HOG3D (Ours)	**7.45**	**2.96**	**37.21**	8.35	**3.44**	**4.19**	11.88	15.62	**13.76**	**4.12**	**10.90**

discriminative weights for the combination of kernels. For dataset, we use TRECVid MED 2011 corpus [44] as our data, due to its large size, realistic content variability, and existing clip-level annotations for 15 different event classes. We compare our approaches (i.e., VFP-EQ-HOG3D and VFP-MI-HOG3D) with the related methods HOG3D, multi-way local pooling using equal kernel weights (MLP-EQ) and MLP using multiple kernel learning (MLP-MKL) [15]. Fig. 10 and Table. V show the evaluation results, which indicates that the our approach achieves better results than the compared methods.

VIII. CONCLUSION

We have presented a framework for videography learning and analysis, and its application for video retrieval, video summarization and videography based feature pooling. The introduced features and data-driven VD learning helps identify characteristic videography among videos from same events. Our experiments show that meaningful summarization and retrieval results can be obtained using videography. The proposed VF-Pooling schema can effectively improve the representation power of features. Both fusion and feature pooling results indicate that videography captures unique aspects of videos and can be jointly used with other features to improve content based video analysis substantially. Our extensive experiments on the challenging TRECVid MED 2011 dataset demonstrate the usefulness of the proposed feature and learning framework. Future work will extend the semantic advantages of videography feature by using it for other high-level content analysis task together with other types of semantic features, such as Action Bank [38]. Also, for videography based feature pooling, more advanced kernel weight learning techniques can be considered such as multiple kernel learning (MKL) [45]. We believe that videography analysis can help solve many more widely-studied video problems.

REFERENCES

[1] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[2] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[4] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: Multi-way local pooling for image recognition," in *Proc. ICCV*, 2011, pp. 2651–2658.

[5] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith, "Scene aligned pooling for complex video recognition," in *Proc. ECCV*, 2012, pp. 688–701.

[6] S. Chakraborty, O. Tickoo, and R. Iyer, "Towards distributed video summarization," in *Proc. 23rd Annu. ACM Conf. Multimedia*, 2015, pp. 883–886.

[7] C. T. Dang and H. Radha, "Heterogeneity image patch index and its application to consumer video summarization," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2704–2718, Jun. 2014.

[8] D. Ding *et al.*, "Beyond audio and video retrieval: Towards multimedia summarization," in *Proc. 2nd ACM Int. Conf. Multimedia Retr.*, 2012, pp. 2:1–2:8.

[9] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3090–3098.

[10] M. A. Hasan, M. Xu, X. He, and C. Xu, "CAMHID: Camera motion histogram descriptor and its application to cinematographic shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1682–1695, Oct. 2014.

[11] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. ECCV*, 2010, pp. 494–507.

[12] R. R. Iyer, S. Parekh, V. Mohandoss, A. Ramsurat, B. Raj, and R. Singh. (2016). "Content-based video indexing and retrieval using corr-LDA." [Online]. Available: https://arxiv.org/abs/1602.08581

[13] W. Jiang, C. Cotton, and A. C. Loui, "Automatic consumer video summarization by audio and visual analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.

[14] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2011, pp. 29:1–29:8.

[15] I. Kim, S. Oh, A. Vahdat, K. Cannons, A. G. Perera, and G. Mori, "Segmental multi-way local pooling for video recognition," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 637–640.

[16] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. BMVC*, 2008, pp. 275-1–275-10.

[17] E. Kraft and T. Brox, "Motion based foreground detection and poselet motion features for action recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 350–365.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008, pp. 1–8.

[19] G. Li, S. Ma, and Y. Han, "Summarization-based video caption via deep neural networks," in *Proc. 23rd Annu. ACM Conf. Multimedia*, 2015, pp. 1191–1194.

[20] K. Li, S. Li, and Y. Fu, "Early classification of ongoing observation," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 310–319.

[21] K. Li, S. Li, and Y. Fu, "Time series modeling for activity prediction," in *Human Activity Recognition and Prediction*. Cham, Switzerland: Springer, 2016, pp. 153–174.

[22] K. Li, S. Oh, A. A. Perera, and Y. Fu, "A videography analysis framework for video retrieval and summarization," in *Proc. BMVC*, 2012, pp. 1–12.

[23] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vis.*, vol. 107, no. 1, pp. 20–39, Mar. 2014.

[24] S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4453–4461.

[25] S. Li, K. Li, and Y. Fu, "Self-taught low-rank coding for visual learning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[26] S. Li, Y. Li, and Y. Fu, "Multi-view time series classification: A discriminative bilinear projection approach," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 989–998.
[27] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,"' in *Proc. CVPR*, 2009, pp. 1996–2003.
[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[29] P. Natarajan *et al.*, "Multimodal feature fusion for robust event detection in Web videos," in *Proc. CVPR*, 2012, pp. 1298–1305.
[30] D. H. Nga and K. Yanai, "Automatic construction of an action video shot database using Web videos," in *Proc. ICCV*, 2011, pp. 527–534.
[31] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. ECCV*, 2010, pp. 392–405.
[32] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[33] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with Fisher vectors on a compact feature set," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1817–1824.
[34] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 540–555.
[35] G. G. Lakshmi Priya and S. Domnic, "Walsh–Hadamard transform kernel-based feature vector for shot boundary detection," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5187–5197, Dec. 2014.
[36] G. B. Rath and A. Makur, "Iterative least squares and compression based estimations for a four-parameter linear global motion model and global motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 7, pp. 1075–1099, Oct. 1999.
[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[38] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1234–1241.
[39] J. Shi and C. Tomasi, "Good features to track," in *Proc. CVPR*, 1994, pp. 593–600.
[40] M. A. Smith, and T. Kanade, "Video skimming for quick browsing based on audio and image characterization," Carnegie Mellon Univ., Pittsburg, PA, USA, Tech. Rep. CMU-CS-95-186, Jul. 1995.
[41] C. Sun and R. Nevatia, "Large-scale Web video event classification by use of Fisher vectors," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 15–22.
[42] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.
[43] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proc. COLING*, vol. 2. 2014, p. 9.
[44] (2011). *Evaluation Plan v3. 0*. [Online]. Available: https://www.nist.gov/sites/default/files/documents/itl/iad/mig/MED11-EvalPlan-V03-20110801a.pdf
[45] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. ICML*, 2009, pp. 1065–1072.
[46] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence—Video to text," in *Proc. ICCV*, 2015, pp. 4534–4542.
[47] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.
[48] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. CVPR*, 2001, pp. I-511–I-518.
[49] H. L. Wang and L. F. Cheong, "Taxonomy of directing semantics for film shot classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 10, pp. 1529–1542, Oct. 2009.
[50] X. Wang, Y.-G. Jiang, Z. Chai, Z. Gu, X. Du, and D. Wang, "Real-time summarization of user-generated videos based on semantic recognition," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 849–852.
[51] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 2742–2746.
[52] A. Whitehead, P. Bose, and R. Laganiere, "Feature based cut detection with automatic threshold selection," in *Proc. Int. Conf. Image Video Retr.*, 2004, pp. 410–418.
[53] J. Yuan *et al.*, "Tsinghua University at TRECVID 2005," in *Proc. TRECVID Workshop*, 2005, pp. 1–17.
[54] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin, "InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval," *IEEE Trans. Multimedia*, vol. 7, no. 4, pp. 648–666, Aug. 2005.
[55] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

Kang Li received the B.S. degree in information and computational science and the M.S. degree in expert system and intelligent control from Northwestern Polytechnical University, China, in 2004 and 2007, respectively, the M.S. degree in computer science and engineering from the State University of New York at Buffalo, Buffalo, in 2011, and the Ph.D. degree in computer engineering from Northeastern University, Boston, MA, USA, in 2015. His research interests include computer vision, applied machine learning, and data mining.

Sheng Li (S'11) received the B.Eng. degree in computer science and engineering and the M.Eng. degree in information security from the Nanjing University of Posts and Telecommunications, China, in 2010 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. He was a Data Scientist Intern with Adobe Research, San Jose, CA, USA, in 2014 and 2015. He has authored 40 papers at leading conferences and journals. His research interests include low-rank matrix recovery, data mining, and machine learning. He received the best paper awards (or nominations) at the SDM 2014, the IEEE ICME 2014, and the IEEE FG 2013. He serves as the Reviewer for several IEEE transactions, and serves as a Program Committee Member of the IJCAI, the AAAI, the IEEE FG, the PAKDD, and the DSAA.

Sangmin Oh received the B.S. degree from Seoul National University in 2003 and the M.S. and Ph.D. degrees from Georgia Tech in 2008 and 2009, respectively. He has authored over 30 papers in the computer vision area. His research has spanned multimedia analysis and retrieval, activity/event recognition from videos, robotic perception, social network analysis, wearable computing, and visualization. His research objective is to present both theoretical and practical routes for challenging real-world problems, where the data is noisy, partially missing, very large, and highly uncertain. He is a Co-Organizer of the tutorial on emerging topics in activity recognition at CVPR'14 and a tutorial on activity recognition at AVSS '12.

Yun Fu (S'07–M'08–SM'11) received the B.Eng. degree in information engineering and the M.Eng. degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, and the M.S. degree in statistics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. He has been an Interdisciplinary Faculty Member affiliated with the College of Engineering and the College of Computer and Information Science, Northeastern University, since 2012. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. His research interests are machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He is a fellow of IAPR, a Lifetime Senior Member of ACM and SPIE, Lifetime Member of the AAAI, the OSA, and the Institute of Mathematical Statistics, a member of the Global Young Academy and the INNS, and a Beckman Graduate Fellow from 2007 to 2008. He received seven Prestigious Young Investigator Awards from the NAE, the ONR, the ARO, the IEEE, the INNS, the UIUC, and the Grainger Foundation, seven Best Paper Awards from the IEEE, the IAPR, the SPIE, and the SIAM, three major Industrial Research Awards from Google, Samsung, and Adobe. He serves as an Associate Editor, the Chair, a PC member, and Reviewer of many top journals and international conferences/ workshops. He is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEANING SYSTEMS.