Bag of Attributes for Video Event Retrieval

Leonardo A. Duarte[1], Otávio A. B. Penatti[2], and Jurandy Almeida[1]

[1]Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil
Email: {leonardo.assuane, jurandy.almeida}@unifesp.br

[2]Advanced Technologies
SAMSUNG Research Institute
13097-160, Campinas, SP – Brazil
Email: o.penatti@samsung.com

***Abstract*—In this paper, we present the Bag-of-Attributes (BoA) model for video representation aiming at video event retrieval. The BoA model is based on a semantic feature space for representing videos, resulting in high-level video feature vectors. For creating a semantic space, i.e., the attribute space, we can train a classifier using a labeled image dataset, obtaining a classification model that can be understood as a high-level codebook. This model is used to map low-level frame vectors into high-level vectors (e.g., classifier probability scores). Then, we apply pooling operations to the frame vectors to create the final bag of attributes for the video. In the BoA representation, each dimension corresponds to one category (or attribute) of the semantic space. Other interesting properties are: compactness, flexibility regarding the classifier, and ability to encode multiple semantic concepts in a single video representation. Our experiments considered the semantic space created by state-of-the-art convolutional neural networks pre-trained on 1000 object categories of ImageNet. Such deep neural networks were used to classify each video frame and then different coding strategies were used to encode the probability distribution from the softmax layer into a frame vector. Next, different pooling strategies were used to combine frame vectors in the BoA representation for a video. Results using BoA were comparable or superior to the baselines in the task of video event retrieval using the EVVE dataset, with the advantage of providing a much more compact representation.**

I. Introduction

The retrieval of videos from specific events, e.g., the wedding of Prince William and Kate Middleton or the riots of 2012 in Barcelona, is a challenging application, as the goal is to retrieve other videos from that event in a database containing lots of different events. This task is even more challenging if we are considering only visual content, i.e., no textual annotations. Different events can occur at the same locations but in different dates, making videos of such events very similar visually. Other challenge is that there can be a large variation in visual aspects, even in the same event. For instance, for the wedding of Prince William and Kate Middleton, there can be videos with close-ups in the people and videos of the location (church, city buildings, etc).

Traditional video descriptors are usually based on low-level features, like textures and local patches [1]–[3], which rarely represent semantic properties. Some more recent approaches aim at including semantics in video representations [4]–[6]. Action Bank [7] is a method to represent videos according to a bank of action detectors. Bag of Scenes [8] considers a dictionary of scenes instead of a dictionary of local patches, and uses it for video geocoding, as the scenes can be representative of places. Works related to semantic signatures [9]–[12] use concept detectors to obtain high-level video representations. Many works are based on obtaining frame-level feature vectors and then aggregating them as a video vector [13]–[19].

Identifying and representing semantics of a video content is one of the most important aspects for video analysis, classification, indexing, and retrieval. If we could have a representation that can encode the multiple elements that appear in a given event in a single feature vector, we could better describe such event and discriminate it from others. Such a representation can be achieved by considering a classifier of high-level concepts in the video. Such concepts could be objects, scenes, locations, and so on.

To achieve such high-level representation for video event retrieval, we present the Bag-of-Attributes (BoA) model. The BoA model is based on a semantic feature space for representing video content, i.e., the attribute space, which can be understood as a high-level visual codebook. This space can be created by training a classifier using a labeled image dataset. Video contents can then be described by applying the learned classifier. The video vector contains the responses of the classifier, in which we have the activations of the semantic concepts that appear in the video. Such representation is a high-level feature vector for the video.

We validated the BoA model for video event retrieval using the EVVE dataset [20]. For obtaining the semantic feature space, we used state-of-the-art convolutional neural networks (CNNs) pre-trained on 1000 object categories of ImageNet. These deep neural networks were used to classify each video frame and then different coding strategies were used to encode the probability distribution from the softmax layer as a high-level frame vector. Next, different pooling strategies were applied over the frame vectors, creating the final video vector (i.e., the bag of attributes). Results point that the BoA model can provide comparable or superior effectiveness to existing

baselines for video event retrieval, with the advantage of BoA generating more compact feature vectors.

The remainder of this paper is organized as follows. Section II introduces some basic concepts and describes related work. Section III presents the BoA model and shows how to apply it for representing video data. Section IV reports the results of our experiments and compares our technique with other methods. Finally, we offer our conclusions and directions for future work in Section V.

II. Background and Related Work

The BoA model is a high-level representation for video event retrieval. Thus, it is related to the areas of visual dictionaries, video representation, and to the task of video event retrieval. In the following sections, we give some background and present related work in each area separately.

A. Visual Dictionaries

Visual dictionaries have been the state-of-the-art representation for many years in visual recognition. Bag of Visual Words, which are representations based on visual dictionaries, have the ability of preserving the discriminating power of local descriptions while pooling those features into a single feature vector [21].

To obtain a Bag-of-Visual-Words (BoVW) representation, the visual dictionary or codebook needs to be created, so the visual content of interest can be represented according to the visual dictionary. Visual dictionaries are created by quantizing a feature space, usually using unsupervised learning approaches and based on a feature space of local features. Such local features are extracted from a training set of images or videos. For extracting local features, one can use interest point detectors or employ dense sampling in a regular grid to obtain local patterns to be described. Then, each local pattern is described by local features, like SIFT and STIP. Those local features are clustered or randomly sampled in order to obtain the visual words of the dictionary. The clustering of the feature space is based solely on the visual appearance and, hence, the visual words themselves carry no semantics [22].

Thereafter, the created dictionary is used to represent the visual content of interest. This is performed by a step usually called *coding*, in which the local features of the content of interest are encoded according to the dictionary. The coding step can employ *hard* or *soft* assignment, for instance. After encoding local features according to the dictionary, a *pooling* step is applied to summarize the assignment values and generate the final feature vector. Popular pooling strategies are *average* and *max* pooling.

As explained above, visual dictionaries are commonly based on unsupervised learning, therefore, having no explicit semantics. The BoA model aims to comprise all the advantages of a visual dictionary model by yet including semantics in the visual words.

B. Video Representation

In the literature, there are many works that obtain high-level representations for videos [4], [6]–[12]. Action Bank [7] is a method to represent videos according to a bank of action detectors. Each video is thus represented by its activations to the action bank. Bag of Scenes (BoS) [8], originally proposed for video geocoding, uses visual codebooks based on whole scenes instead of local patches (e.g., corners or edges). Such scenes represent places of interest, therefore, for geocoding, the BoS vector works as a place activation vector.

Some existing methods are based on obtaining frame-level feature vectors and then aggregating them as a video vector [11], [13]–[19]. Such methods, which also includes the BoA model, can make use of the great variety of works that aim to obtain semantic representations for images [5], [22], [23]. For video retrieval, Mazloom et al. [11] use average pooling to aggregate frame vectors obtained by concept detectors. Jiang et al. [13] make use of Object Bank [22] and Classemes [23] to obtain the high-level frame vectors, which are then aggregated as the video vectors. Some of such works [17]–[19] make use of CNNs for extracting frame-level features, however, they allow the use of CNN layers that do not have explicit semantics (e.g., fully-connected or even convolutional layers). Therefore, they are conceptually different from our work. For BoA, the only CNN layer that makes sense is the last layer, which is the only one with explicit semantics.

Except for BoS, all the other works mentioned above do not deal with high-level video representations in terms of high-level visual codebooks as we do in the BoA model. Unlike BoS which uses no information from labels or from classifiers, BoA explicitly uses the labels given by supervised classifiers as feature vectors for video frames, which are then aggregated in the BoA vector.

C. Video Event Retrieval

A comprehensive review of event-based media processing and analysis methods can be found in [24]. Most of existing research works has focused on video event detection or video event recounting. In this work, we are interested in video event retrieval. Although related, there are substancial differences between such tasks [25].

In [20], frame-based features are encoded into a temporal representation for videos. Initially, dense SIFT descriptors were extracted from video frames and then PCA was used to reduce them to 32 dimensions. Next, the reduced SIFT descriptors of each frame were encoded into a frame vector using MultiVLAD. After that, two different approaches for encoding frame vectors into a video representation were evaluated. The former, called Mean-MultiVLAD (MMV), aggregates the frame vectors over the entire video using average pooling. The latter, called Circulant Temporal Encoding (CTE), creates a temporal representation that jointly encodes the frame vectors in the frequency domain. In the experiments, both methods are combined, creating the MMV+CTE representation.

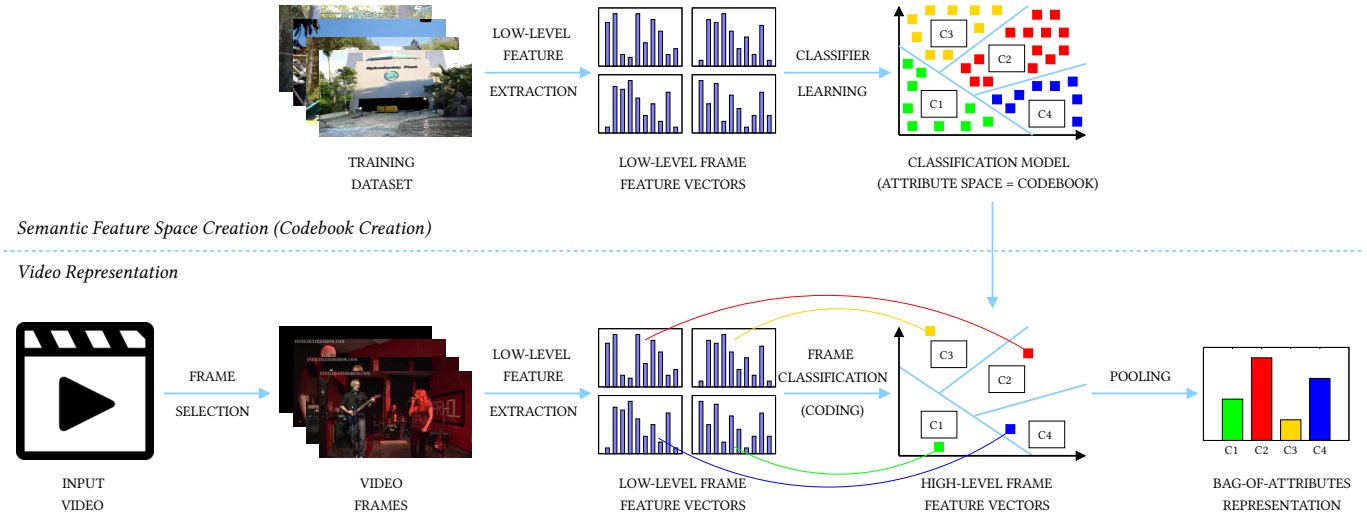

Fig. 1. Overview of the Bag-of-Attributes model. On top, we show how to obtain a semantic feature space for then using it to represent videos. At the bottom, we show how to map video content into this semantic space. As the process is similar to the creation and use of visual codebooks, we also show the names that are commonly used in that scenario. CN represent the categories (classes or attributes) in the training dataset.

In [26], different hyper-pooling strategies are proposed to encode frame-level feature vectors into a video-level feature vector. The underlying idea relies on two stages: (i) hashing, where frame vectors are distributed into different cells; and (ii) encoding, where frame vectors in each cell are aggregated by accumulating residual vectors. Each frame was represented using the same approach adopted in [20], as described above. Four different hashing functions were evaluated: k-means, partial k-means (PKM), sign of stable components (SSC) e KD-Tree. The best result was obtained with SSC.

III. Bag of Attributes

In this section, we present the Bag-of-Attributes (BoA) model for video representation. The main objective of the BoA model is to represent videos in a feature space with semantic information, resulting in a high-level representation [8], [22]. For that, we basically need to have a semantic feature space and a mapping function from the original video space to this new space. The steps involved in the BoA model are presented in Figure 1.

In the BoA model, we obtain the semantic feature space by training a supervised classifier based on a labeled image dataset. The learned classifier, thus incorporates semantic information learned from the dataset. We call each label of the learning set as an *attribute*, aiming at referring to elements containing semantics.

For mapping or coding the video properties in this semantic (high-level) feature space, we start by classifying each frame of the input video with the learned classifier. Therefore, each frame is represented by its classification results, creating a high-level feature vector. Such results can be simply the class label given by the classifier or the whole probability vector, containing the probabilities of that frame in relation to every attribute of the learned classifier. Then, after having a high-level feature vector for each video frame, we generate the final video representation by computing some statistical measure over the frame vectors.

An obvious but important remark about the low-level feature extraction from video frames: in both stages (creation of the semantic space and video representation, i.e., top and bottom parts of Figure 1), the low-level feature space must be the same. For instance, if the classifier was trained with frames represented by color histograms, the classifier, of course, can only be applied over color histograms. Therefore, in this example, the frames of the video to be represented by BoA must have color histograms as low-level feature vectors.

We can easily map the steps in the BoA model to the steps involved in the context of visual dictionaries and bags of visual words. The learned classifier can be seen as the *codebook*: each visual word is an attribute, i.e., a region in the classification space. The process of classifying each frame with the learned classifier can be seen as the *coding* step (visual word assignment). If in this step we consider only the classifier final attribution, i.e., class label for the frame, we have something similar to hard assignment. If we consider the classifier probability vector, we have something similar to soft assignment [27], [28]. Then, the final step of summarizing the frame representations can be seen as the *pooling* step, which can be implemented by summing, averaging or considering the maximum probability score among frames for each class [21].

Some interesting properties of the BoA representation are: (i) one dimension for each semantic concept, (ii) compactness (dimensionality equal to the number of classes in the learned classifier), (iii) flexibility to use any kind of classifier for creating the semantic feature space, and (iv) ability to encode multiple semantic concepts in a single representation. The last property can be understood if we consider that in the pooling operation we keep probability scores of the multiple

TABLE I
EVVE EVENTS LIST. THE DATASET HAS A TOTAL OF 620 QUERY VIDEOS AND 2,375 DATABASE VIDEOS DIVIDED INTO 13 EVENTS. Q REFERS TO THE NUMBER OF QUERIES, DB+ AND DB- ARE THE NUMBERS OF POSITIVE AND NEGATIVE VIDEOS IN THE DATABASE, RESPECTIVELY.

ID	Event name	Q	Db+	Db-
1	Austerity riots in Barcelona, 2012	13	27	122
2	Concert of Die toten Hosen, Rock am Ring, 2012	32	64	143
3	Arrest of Dominique Strauss-Kahn	9	19	60
4	Egyptian revolution: Tahrir Square demonstrations	36	72	27
5	Concert of Johnny Hallyday stade de France, 2012	87	174	227
6	Wedding of Prince William and Kate Middleton	44	88	100
7	Bomb attack in the main square of Marrakech, 2011	4	10	100
8	Concert of Madonna in Rome, 2012	51	104	67
9	Presidential victory speech of Barack Obama 2008	14	29	56
10	Concert of Shakira in Kiev 2011	19	39	135
11	Eruption of Strokkur geyser in Iceland	215	431	67
12	Major autumn flood in Thailand, 2011	73	148	9
13	Jurassic Park ride in Universal Studios theme park	23	47	10
All	>>>	**620**	**1252**	**1123**

classes activated over the video frames. For instance, if our attribute space is based on objects (like the object categories of ImageNet [29]), each frame will be classified considering the presence or not of such objects in the frame. The final video vector will then contain information of the objects that appear along the video. The BoA representation can be generalized to many other scenarios, which depend only on the attribute space to be considered. Other possible examples could be by considering classifiers trained to categorize scenes, faces, plants, vehicles, actions, etc.

For implementing the BoA model, different approaches can be used. For video frame selection, techniques like sampling at fixed-time intervals or summarization methods [30], [31] could be employed. For creating the attribute classifier (i.e., the codebook), which is one of the key steps in the BoA model, one can learn the classifier in the dataset which better represents the contents of interest. Other option is to employ existing pre-trained classifiers, like the state-of-the-art classifiers based on CNNs [32]–[38], which were trained on 1000 object categories of ImageNet dataset [29]. In this case, the low-level feature extraction step of the BoA model is also performed by the deep neural networks, as CNNs integrate both feature extraction and classification abilities. Therefore, video frames can be directly used as input for the CNN and its output will be the frame high-level feature vectors, considering the use of the final classification layer, which is commonly a soft-max layer.

By considering a semantic space of object categories, videos will be represented according to the existence or not of objects along their frames. This information can help in discriminating videos from different categories. For instance, if we have presence of object categories like weapon, knife, blast and/or fire, the video is possibly related to violent events.

IV. EXPERIMENTS AND RESULTS

The BoA model was validated for video event retrieval. Section IV-A describes the dataset and the protocol adopted in the experimental evaluation. Section IV-B discusses the impact of parameters and options for generating the BoA representation. Finally, a comparison with the methods discussed in Section II-C is presented in Section IV-C.

A. The EVVE Dataset

Experiments were conducted on the EVVE (EVent VidEo) dataset[1]: an event retrieval benchmark introduced by Revaud et al. [20]. This dataset is composed of 2,995 videos (166 hours) collected from YouTube[2]. Those videos are distributed among 13 event categories and are divided into a query set of 620 (20%) videos and a reference collection of 2,375 (80%) videos. Each event is treated as an independent subset containing some specific videos to be used as queries and the rest to be used as database for retrieval, as shown in Table I. It is a challenging benchmark since the events are localized in both time and space. For instance, event 1 refers to the great riots and strikes that happened in the end of March 2012 at Barcelona, Spain, however, in the database, there are a lot of videos from different strikes and riots around the world.

EVVE uses a standard retrieval protocol: a query video is submitted to the system which returns a ranked list of similar videos. Then, we evaluate the average precision (AP) of each query and compute the mean average precision (mAP) per event. The overall performance is assessed by the average of the mAPs (avg-mAP) obtained for all the events.

B. Impact of the Parameters

Our experiments followed the official experimental protocol created by Revaud et al. [20]. Initially, each video in the dataset was represented by a BoA vector. With the BoA of each video, a given query video was used to retrieve the rest database videos, which were ranked according to the cosine similarity between their BoAs. Finally, we used the dataset official tool to evaluate the retrieval results[3]. It is important to emphasize that we analyze only the visual content, ignoring audio information and textual metadata.

[1]http://pascal.inrialpes.fr/data/evve/ (As of June, 2018).
[2]http://www.youtube.com (As of June, 2018).
[3]http://pascal.inrialpes.fr/data/evve/eval_evve.py (As of June, 2018).

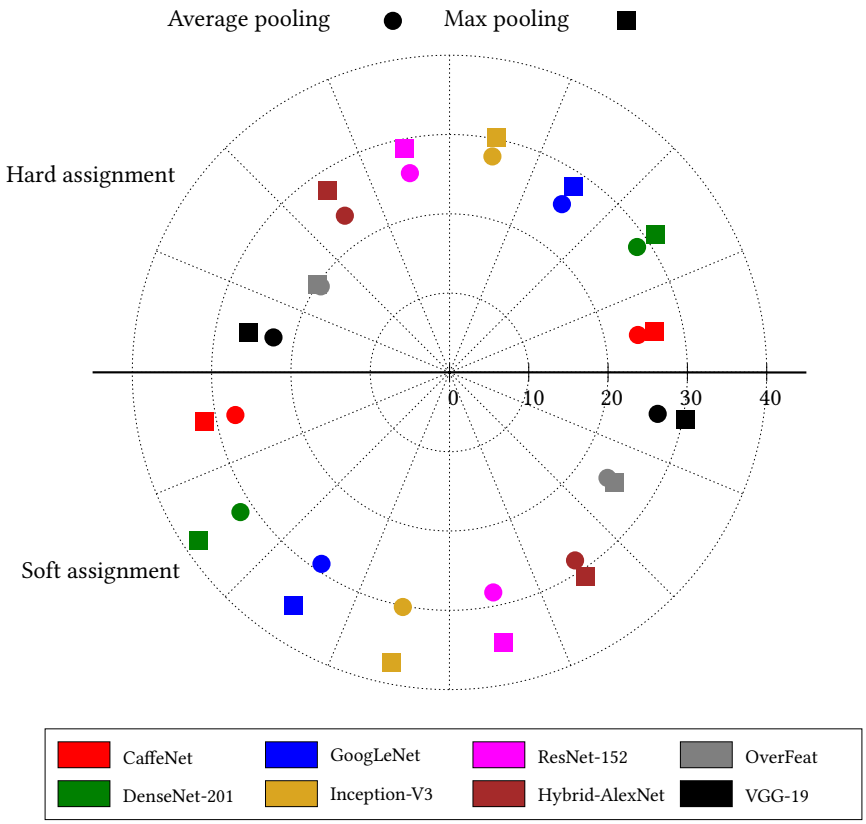

Fig. 2. Results obtained by different configurations of BoA. The sectors refer to the CNNs used as attribute classifier. Each semicircle is related to a coding strategy. Different symbols were associated to the pooling strategies. The radius indicates the avg-mAP for each parameter setting. The farther from the center, the higher the avg-mAP. We can note that the best BoA configuration is based on DenseNet-201 with soft assignment and max pooling.

In this paper, video frames were selected using the well-known FFmpeg tool[4] in a sampling rate of one frame per second. For obtaining the semantic feature space, we considered eight different CNN architectures: CaffeNet [34], DenseNet-201 [33], GoogLeNet [36], Inception-V3 [37], Microsoft ResNet-152 [32], MIT Places Hybrid-AlexNet[5] [39], OverFeat[6] [35], and VGG ILSVRC 19 Layers [38]. We used all the CNNs trained on 1000 object categories of ImageNet [29], except for MIT Places Hybrid-AlexNet, which considered both objects and scene categories, more specifically, 978 object categories of ImageNet and 205 scene categories of Places Database [39]. In this way, we evaluated not only the result of different classifiers, but also the effect of different types of attributes. Each frame was classified by such CNNs and then represented by a feature vector corresponding to the probability distribution from their softmax layer. To obtain the BoA representation of each video, we evaluated the use of hard and soft assignment (i.e., only the predicted class or the whole probability vector, respectively) as well as average and max pooling.

First, we conducted experiments aiming at determining the best parameter setting for the BoA model. Different options were evaluated for each of the steps involved in creating the BoA representation. Table II summarizes the parameters and options evaluated in our experiments.

TABLE II
THE PARAMETERS AND OPTIONS EVALUATED FOR GENERATING THE BOA REPRESENTATION.

Parameter	Options
Attribute classifier	CaffeNet, DenseNet-201, GoogleLeNet, Inception-V3, ResNet-152, Hybrid-AlexNet, OverFeat, VGG-19
Coding technique	hard, soft
Pooling technique	average, max

Figure 2 presents a radar plot showing the results obtained for different configurations of the BoA model. This chart is composed of three parts: sector, points, and radius. Each sector indicates the results obtained by using a given CNN as the attribute classifier. In order to make the comparison easier, all the sectors from a same coding strategy were grouped together: the results at the top semicircle refers to the hard assignment (i.e., only the predicted class) whereas those related to the soft assigment (i.e., the whole probability vector) are presented at the bottom semicircle. The points in each sector were associated to different symbols and each of them represents a

[4]http://www.ffmpeg.org/ (As of June, 2018).
[5]places.csail.mit.edu/downloadCNN.html (As of June, 2018)
[6]http://cilvr.nyu.edu/doku.php?id=software:overfeat:start (As of June, 2018).

different pooling strategy. Finally, the radius denotes the avg-mAP. The farther a point is from the center, the better the results (i.e., the higher the avg-mAP).

In general, the soft assignment (bottom semicircle) yielded better results than the hard assignment (top semicircle) for performing the coding step of BoA. The soft assignment takes into account the semantic ambiguity, i.e., a same frame may activate two or more attributes, which is neglected in the hard assignment, where just one attribute is activated for each frame, thus discarding relevant information about its visual content. With respect to the pooling strategies, max pooling (squares) yielded the highest avg-mAP scores, regardless of the attribute classifier or coding strategy used in the BoA model. For event retrieval, no significant difference was observed in the use of only objects (i.e., CaffeNet, DenseNet-201, GoogleLeNet, Inception-V3, ResNet-152, OverFeat, VGG-19) or both objects and scenes (i.e., Hybrid-AlexNet). The best results were observed using DenseNet-201 as the attribute classifier, reaching an avg-mAP of more than 38%.

In the next experiments, when we refer to BoA, we mean the DenseNet-201 was used for classifying video frames, soft assignment (i.e., the whole probability vector) was used for coding frame-level feature vectors, and max pooling was used to summarize them in a video-level feature vector. With such representation, the BoA vector contains a summary of the objects present in a video.

C. Comparison with Baselines

We compared the results obtained by BoA with those reported by Revaud et al. [20] for three baselines: Mean-MultiVLAD (MMV), CTE (Circulant Temporal Encoding) and a combination of both methods, known as MMV+CTE. Also, we considered the results reported by Douze et al. [26] for the variations of MMV with the following hyper-pooling functions: k-means, partial k-means (PKM), sign of stable components (SSC), KD-Tree and Fisher Vectors (FV).

In Figure 3, we compare the BoA representation and the baseline methods with respect to the avg-mAP. As we can observe, the BoA model yielded better results than all the baseline methods.

The results where also compared by event, as shown in Table III. Notice that BoA performed better than the baseline methods for most of the events (7 out of 13). For some events, the difference in favor of the BoA model is very large. For instance, for event 12 (*"Major autumn flood in Thailand, 2011"*), the best baseline (MMV+CTE) achieves 37.1% of avg-mAP while BoA obtains 57.1%.

We also performed paired t-tests to verify the statistical significance of the results. For that, the confidence intervals for the differences between paired averages (mAP) of each category were computed to compare every pair of approaches. If the confidence interval includes zero, the difference is not significant at that confidence level. If the confidence interval does not include zero, then the sign of the difference indicates which alternative is better.

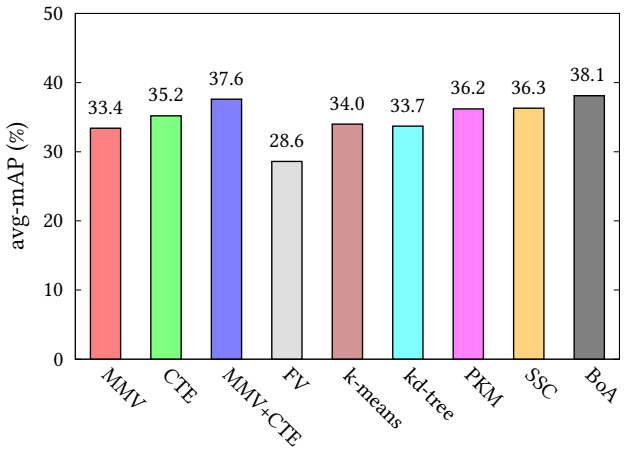

Fig. 3. Performance of different methods for event retrieval on EVVE dataset. The BoA representation performed better than all the baselines.

TABLE III
RETRIEVAL PERFORMANCE (MAP) PER EVENT ON EVVE DATASET.

Event ID	MMV	CTE	MMV+CTE	BoA
1	23.9	13.9	24.6	**32.4**
2	19.9	16.6	20.2	**20.8**
3	8.7	12.8	11.1	**13.8**
4	12.6	10.8	**13.2**	12.6
5	23.4	26.2	26.0	**29.8**
6	33.8	41.3	39.4	**49.1**
7	12.4	**25.2**	21.2	14.9
8	25.4	25.7	**28.1**	26.7
9	53.1	**80.3**	69.4	52.3
10	45.5	40.9	**48.6**	30.0
11	77.3	71.4	77.4	**86.4**
12	36.6	29.7	37.1	**57.1**
13	60.4	69.3	**71.9**	69.8
avg-mAP	33.4	35.2	37.6	**38.1**

Figure 4 presents the confidence intervals (for $\alpha = 0.05$) of the differences between BoA and the baseline methods for the mAP measures. As we can observe, the confidence intervals for BoA and MMV are positive, indicating that BoA outperformed MMV. On the other hand, the confidence intervals for BoA and CTE include zero and, hence, the differences between such approaches are not significant at that confidence level.

We believe that BoA outperformed MMV because the BoA representation carries semantic information. The explicit enconding of object occurrences in the BoA representations may have created a better feature space separability for videos of different events. On the other hand, BoA does not include temporal information and we think such feature is important to recognize some types of events. Unlike BoA, CTE is a more complex method that exploits the temporal information of a video. For that reason, CTE is better than MMV and BoA for describing events composed by repeatable small sequences, like the event 9 (*"Presidential victory speech of Barack Obama 2008"*).

Similar to the pooling step of the BoA model, the reasoning behind CTE is to summarize a set of frame vectors in a

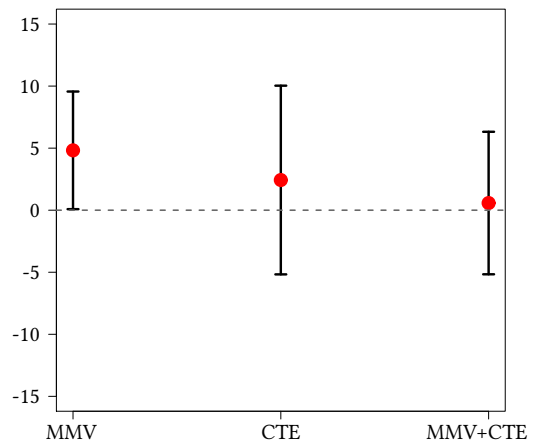

Fig. 4. Paired t-test comparing BoA and the baselines. BoA outperformed MMV with statistical significance (intervals above zero), while had no significant difference to CTE and MMV+CTE (intervals cross zero), although BoA presented higher avg-mAP.

single video vector. However, different from average and max pooling, CTE is able to encode the temporal order of a frame sequence. Therefore, in the same sense of MMV+CTE, BoA and CTE are complementary and could be combined in order to obtain a more discriminant representation. For that, we could use CTE instead of max pooling to summarize the frame vectors in the BoA model.

The key advantage of the BoA model is its computational efficiency in terms of space occupation and similarity computation time. Since the time required to compute the similarity among videos is hardware dependent (i.e., with faster hardware the computational speed increases and the running time decreases) and the source codes of all the baselines methods are not available, it is challenging to perform a fair comparison of performance in relation to BoA. Table IV compares the BoA model with the baseline methods with respect to the computational complexity for similarity computation and space requirements. In this way, we can investigate the relative difference of performance among such approaches.

TABLE IV
THE COMPUTATIONAL COMPLEXITY FOR SIMILARITY COMPUTATION AND SPACE REQUIREMENTS OF THE DIFFERENT APPROACHES IN TERMS OF THE NUMBER n OF FRAMES IN A VIDEO[7].

Method	*Descriptor Size*	*Similarity Computation*	*Space Requirements*
MMV	512	$O(1)$	$O(1)$
CTE[7]	$p \times \beta \times n$	$\Omega(n)$	$\Omega(n)$
MMV+CTE[7]	$p \times \beta \times n$	$\Omega(n)$	$\Omega(n)$
FV	16384	$O(1)$	$O(1)$
k-means	16384	$O(1)$	$O(1)$
kd-tree	16384	$O(1)$	$O(1)$
PKM	16384	$O(1)$	$O(1)$
SSC	16384	$O(1)$	$O(1)$
BoA	**1000**	$\boldsymbol{O(1)}$	$\boldsymbol{O(1)}$

[7]CTE relies on two parameters: the number p of PQ sub-quantizers and the frame description rate β. For details regarding CTE, please, refer to [20], [25].

As we can observe, the time complexities for both similarity computation and space occupation of CTE and MMV+CTE grows at least linear with the video length. For all the other methods, the computational costs are constant and, hence, they are independent of the video duration. Although MMV is the most compact (i.e., only 512 dimensions), its performance was one of the worst, yielding the second lowest avg-mAP of 33.4% (see Figure 3). Clearly, BoA is much more efficient than the baselines, since it is the second most compact (i.e., less than double of MMV), but the most effective, yielding the highest avg-mAP of 38.1%.

V. CONCLUSIONS

We presented a semantic video representation for video event retrieval, named Bag of Attributes (BoA). In this model, videos are represented in a high-level feature space, which comprises the classification space defined by a supervised image classifier. In such space, each region corresponds to a semantic concept. To represent video content in this space, we start by classifying each video frame with the learned classifier, resulting in a high-level feature vector for each frame (e.g., classifier probability scores). Then, frame vectors are summarized by pooling operations to generate the final video vector, creating the BoA representation.

The main properties of the BoA representation are: each vector dimension corresponds to one semantic concept, compactness, flexibility regarding the learned classifier, and ability to encode multiple semantic concepts in a single vector.

To validate the BoA model for video event retrieval, we conducted experiments on the EVVE dataset. Our implementation of BoA considered the semantic feature space created by state-of-the-art CNNs pre-trained on 1000 object categories of ImageNet. Such CNNs were used to classify video frames. We evaluate the impact of different coding strategies used to encode the probability distribution from the softmax layer as high-level frame vectors. Also, different pooling strategies were tested for summarizing the frame vectors in a final video vector (i.e., the bag of attributes). The results demonstrated that BoA performs similar to or better than the baselines, but using a much more compact representation. We believe that the ability to encode multiple concepts in the BoA representation could improve discriminating between events.

As future work, we plan to evaluate other semantic spaces created by classifiers based on CNNs (e.g., Xception [40] and ResNeXt [41]). In addition, we intend to evaluate different strategies for including temporal information in the representation space (e.g., using recurrent neural networks). We also consider to evaluate classifiers trained on non-object categories, like scenes, for instance. The evaluation of the BoA model in other applications besides event retrieval is also a possible future work.

ACKNOWLEDGMENT

We thank the São Paulo Research Foundation - FAPESP (grant 2016/06441-7) and the Brazilian National Council for Scientific and Technological Development - CNPq

(grants 423228/2016-1 and 313122/2017-2) for funding. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] J. Almeida, N. J. Leite, and R. S. Torres, "Comparison of video sequences with histograms of motion patterns," in *IEEE International Conference on Image Processing (ICIP'11)*, 2011, pp. 3673–3676.

[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2011, pp. 3169–3176.

[3] G. Willems, T. Tuytelaars, and L. J. van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European Conference on Computer Vision (ECCV'10)*, 2008, pp. 650–663.

[4] J. Dalton, J. Allan, and P. Mirajkar, "Zero-shot video retrieval using content and concepts," in *ACM International Conference on Information and Knowledge Management (CIKM'13)*, 2013, pp. 1857–1860.

[5] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol. 108, no. 1, pp. 59–81, 2014.

[6] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan, "Zero-shot event detection using multi-modal fusion of weakly supervised concepts," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'14)*, 2014, pp. 2665–2672.

[7] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'11)*, 2012, pp. 1234–1241.

[8] O. A. B. Penatti, L. T. Li, J. Almeida, and R. da S. Torres, "A visual approach for video geocoding using bag-of-scenes," in *ACM International Conference on Multimedia Retrieval (ICMR'12)*, 2012, pp. 1–8.

[9] A. Agharwal, R. Kovvuri, R. Nevatia, and C. G. M. Snoek, "Tag-based video retrieval by embedding semantic content in a continuous word space," in *IEEE Winter Conference on Applications of Computer Vision (WACV'16)*, 2016, pp. 1–8.

[10] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek, "Recommendations for video event recognition using concept vocabularies," in *ACM International Conference on Multimedia Retrieval (ICMR'13)*, 2013, pp. 89–96.

[11] M. Mazloom, A. Habibian, and C. G. M. Snoek, "Querying for video events by semantic signatures from few examples," in *ACM International Conference on Multimedia (ACM-MM'13)*, 2013, pp. 609–612.

[12] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, 2012.

[13] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," in *AAAI Conference on Artificial Intelligence (AAAI'13)*, 2013, pp. 1113–1119.

[14] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann, "Complex event detection via multi-source video attributes," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'13)*, 2013, pp. 2627–2633.

[15] J. Y.-H. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 4694–4702.

[16] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *ACM International Conference on Multimedia (ACM-MM'15)*, 2015, pp. 471–480.

[17] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 1798–1807.

[18] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *ACM International Conference on Multimedia (ACM-MM'15)*, 2015, pp. 461–470.

[19] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting image-trained CNN architectures for unconstrained video classification," in *British Machine Vision Conference (BMVC'15)*, 2015, pp. 60.1–60.13.

[20] J. Revaud, M. Douze, C. Schmid, and H. Jegou, "Event retrieval in large video collections with circulant temporal encoding," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'13)*, 2013, pp. 2459–2466.

[21] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'10)*, 2010, pp. 2559–2566.

[22] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Conference on Neural Information Processing Systems (NIPS'10)*, 2010, pp. 1378–1386.

[23] A. Bergamo and L. Torresani, "Classemes and other classifier-based features for efficient object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1988–2001, 2014.

[24] C. Tzelepis, Z. Ma, V. Mezaris, B. Ionescu, I. Kompatsiaris, G. Boato, N. Sebe, and S. Yan, "Event-based media processing and analysis: A survey of the literature," *Image Vision Computing*, vol. 53, pp. 3–19, 2016.

[25] M. Douze, J. Revaud, J. J. Verbeek, H. Jégou, and C. Schmid, "Circulant temporal encoding for video retrieval and temporal alignment," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 291–306, 2016.

[26] M. Douze, J. Revaud, C. Schmid, and H. Jegou, "Stable hyper-pooling and query expansion for event detection," in *IEEE International Conference on Computer Vision (ICCV'13)*, 2013, pp. 1825–1832.

[27] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[28] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *IEEE International Conference on Computer Vision (ICCV'11)*, 2011, pp. 2486–2493.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[30] J. Almeida, R. da S. Torres, and N. J. Leite, "Rapid video summarization on compressed video," in *IEEE International Symposium on Multimedia (ISM'10)*, 2010, pp. 113–120.

[31] J. Almeida, N. J. Leite, and R. da S. Torres, "VISON: VIdeo Summarization for ONline applications," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, pp. 770–778.

[33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 2261–2269.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'15)*, 2015, pp. 1–9.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'16)*, 2016, pp. 2818–2826.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Conference on Neural Information Processing Systems (NIPS'14)*, 2014, pp. 487–495.

[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 1800–1807.

[41] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 5987–5995.