Bag of Genres for Video Retrieval

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

[1]Institute of Science and Technology
Federal University of 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*—Often, videos are composed of multiple concepts or even genres. For instance, news videos may contain sports, action, nature, etc. Therefore, encoding the distribution of such concepts/genres in a compact and effective representation is a challenging task. In this sense, we propose the Bag of Genres representation, which is based on a visual dictionary defined by a genre classifier. Each visual word corresponds to a region in the classification space. The Bag of Genres video vector contains a summary of the activations of each genre in the video content. We evaluate the proposed method for video genre retrieval using the dataset of MediaEval Tagging Task of 2012 and for video event retrieval using the EVVE dataset. Results show that the proposed method achieves results comparable or superior to state-of-the-art methods, with the advantage of providing a much more compact representation than existing features.**

***Keywords*-video retrieval; video representation; visual dictionaries; semantics**

I. Introduction

The retrieval of videos by content is a challenging application, as videos may be composed of visually different excerpts. For instance, a news video can comprise multiple categories, like sports, documentary, health, and others. A video retrieval system aiming at retrieving videos with similar content should be aware of such property in order to obtain better results.

In this paper, we focus on video retrieval based only on visual information. No tags or textual descriptions are considered. One important step in this scenario is feature extraction from videos. There are mainly two kinds of feature descriptors for videos: descriptors that consider motion and descriptors based on isolated frames. Motion-based descriptors usually obtain space-time interest points and extract histograms of those local points or obtain histogram of motion patterns [1]. Descriptors based on isolated frames are usually derived from image feature extraction. Frames are represented individually and then a pooling function can be used to obtain the video feature vector. The advantage of the first kind of descriptors is obviously the encoding of transitions between frames. The advantage of the second kind is the possibility to use the large number of descriptors already proposed for image representation.

Regardless of motion, many of the state-of-the-art solutions for feature extraction are based on visual dictionaries. Such dictionaries are commonly based on local patches, which are semantically poor. Therefore, both kinds of descriptors usually present the same property: the video feature vector has few semantics from the human perspective.

In this paper, we present a novel approach for video representation, called Bag of Genres (BoG). The proposed method is based on dictionaries of genres created from genre classifiers. Each visual word in the BoG model is a genre-labeled region of the classification space defined by the classifier's model. The main advantage of the BoG model are the following: (i) each visual word explicitly contains semantics, which was learned from the labeled data by the genre classifier; (ii) the video representation corresponds to an activation vector of its contents to each of the genres in the dictionary, thus having one dimension for each genre; and (iii) compact representation, directly related to the number of genres in the dictionary.

We validated the BoG model for video genre retrieval and for video event retrieval. In the first case, we used the dataset of MediaEval Tagging Task of 2012. We evaluate the importance of the genre classifier in the model as well as the quality of the BoG representation. Although the genre classifier has low accuracy, the BoG model worked well in the experiments. The results are comparable to the existing baselines, but BoG is much more compact. In the second case, we used our best BoG representation to retrieve videos by event on the EVVE dataset. The results in this dataset indicate that the BoG approach outperforms state-of-the-art methods.

The remainder of the paper is organized as follows. Section II presents related work. Section III explains the proposed BoG model. Section IV shows experiments and results, and Section V concludes the work indicating possible future work.

II. Related Work

In this section, we describe related work focusing on works that are based on visual dictionaries and works that aim at including semantics in the representation.

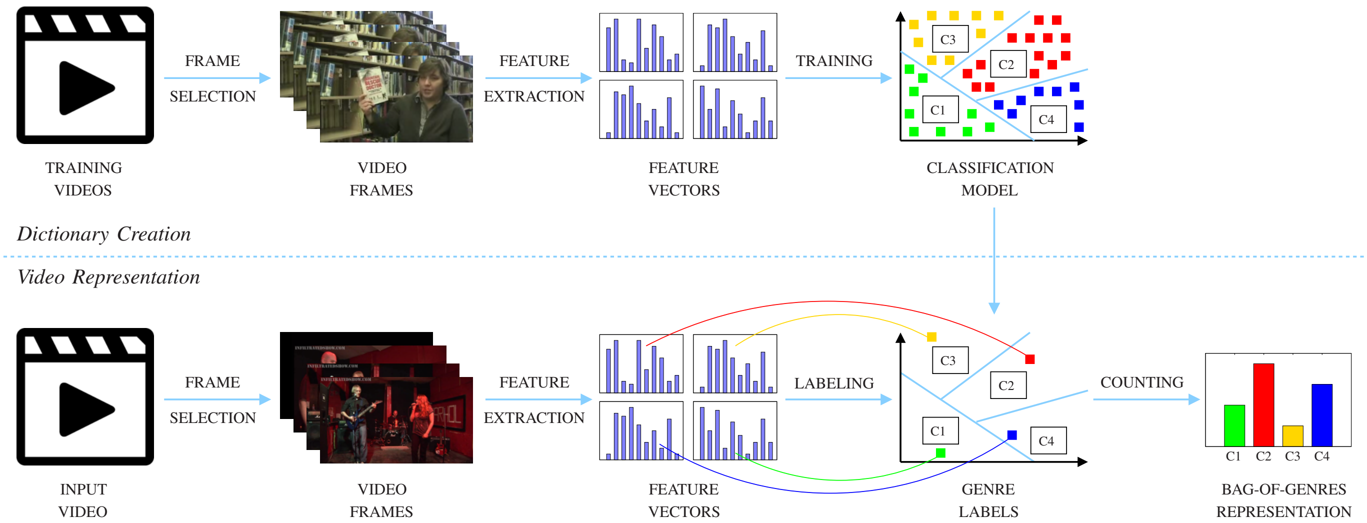


Fig. 1. An overview of the Bag-of-Genres model.

Many solutions exist in the literature aiming at including semantics in the representation. There are techniques in which an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors [2], which could be seen as a dictionary of objects. Poselets have also been used similarly to a dictionary of poses for recognizing people poses [3]. Labeled local patches have also been used for having a dictionary with more semantics [4]. Boureau et al. [5] also present a way to supervise the dictionary creation. Other approaches can also be considered as related to the intention of having dictionaries with more meaningful visual words [6]–[9]

The approach proposed here is closely related to the Bag-of-Scenes (BoS) model [10], in which the video feature vector is an activation vector of scenes. As scenes are more semantically meaningful than local patches, the BoS feature space is semantically richer. Each dimension in the BoS space corresponds to a semantic concept.

The main novelty of BoG in relation to previous works, specially BoS, is that we use a genre classifier as visual dictionary. In the BoS model, the visual dictionary is based directly on the feature vectors of the scenes. The advantages of using a classifier is that it better delineates the frontiers among visual words and tends to be more robust to feature dimensionality. Another advantage is the compact BoG vector, as its dimensionality directly corresponds to the number of genres in the problem.

III. Bag of Genres

In this section, we describe the Bag-of-Genres (BoG) model for video representation. This model is based on a dictionary of genres, in which each visual word corresponds to a decision region of the classification model defined by a genre classifier. Thus, each video is represented by a vector of activations of its frames to each of the genres in the dictionary.

An interesting property of the BoG model is that it relies on elements that have more semantics according to the human perception. Traditional dictionaries based on local features, like SIFT or STIP, are composed of visual words which carry no semantic information, like corners and edges [2]. In the BoG model, as the visual words are genre-labeled regions of the classification space, the activation vector has one dimension for each genre, making it simple to analyze the presence or absence of each genre into a video. Another important aspect of using a genre classifier to encode visual features, is that the classifier better delineates the feature space and classifiers (e.g., Support Vector Machines) deal well with high dimensional spaces.

Figure 1 shows a flowchart of the stages involved in representing video content using the BoG model. On top, we show how the visual dictionary is created. At the bottom, we show how this codebook is used to represent video content.

The creation of the visual dictionary is performed as follows. Given a set of *training videos* with known genre labels, we first discard a lot of redundant information, taking only a subset of *video frames*. Techniques like sampling at fixed-time intervals or summarization methods [11], [12]are examples of possibilities for *frame selection*. In this paper, frames were selected using the well-known FFmpeg tool[1] in a sampling rate of one frame per second. After that, we perform the *feature extraction* from each of the selected frames in order to encode their visual content into *feature vectors*. Such features can be any, like for instance, color histograms, GIST, bags of quantized SIFT features, or even features extracted from deep convolutional neural networks [13]. Then, those feature vectors and their associated genre labels are used as input for *training* a genre classifier. The obtained *classification model* represents the dictionary of genres used for representing videos.

After creating the visual dictionary, we should represent videos according to the dictionary space. Given an *input video*, we initially apply *frame selection* and *feature extraction* from

[1]http://www.ffmpeg.org/ (As of May 2016).

each frame. After that, the feature vectors of each frame must be coded according to the dictionary of genres. Each feature vector is classified by the genre classifier, which predicts a *genre label* for the frame. The *labeling* process is analogous to the *coding* step of traditional visual dictionaries [14]. Finally, a normalized frequency histogram is obtained by *counting* the occurrences of each of the genre labels, forming the *bag-of-genres representation* for the input video. Such step can be understood as *pooling* the frame genres [5].

The dimensionality of the bag-of-genres feature space is directly related to the number of genres used for training the genre classifier during the dictionary creation. Therefore, as in many applications the number of genres is small, the bag of genres is usually more compact than existing solutions.

IV. Experiments and Results

We evaluate the BoG model on two challenging tasks: for video genre retrieval, using the dataset of MediaEval Tagging Task of 2012; and for video event retrieval, using the EVVE dataset. In the following subsections, we report and discuss the obtained results.

A. Video Genre Retrieval

Experiments were conducted on a benchmarking dataset provided by the MediaEval 2012 organizers for the Genre Tagging Task [15]. The dataset is composed of 14,838 videos (3,288 hours) collected from the blip.tv[2] and is divided into a training set of 5,288 videos (36%) and a test set of 9,550 videos (64%). Those videos are distributed among 26 video genre categories assigned by the blip.tv media platform, namely (the numbers in brackets are the total number of videos): art (530), autos and vehicles (21), business (281), citizen journalism (401), comedy (515), conferences and other events (247), documentary (353), educational (957), food and drink (261), gaming (401), health (268), literature (222), movies and television (868), music and entertainment (1148), personal or auto-biographical (165), politics (1107), religion (868), school and education (171), sports (672), technology (1343), environment (188), mainstream media (324), travel (175), video blogging (887), web development (116), and default category (2349, which comprises videos that cannot be assigned to any of the previous categories). The main challenge of this collection is the high diversity of genres, as well as the high variety of visual contents within each genre category [16], [17].

After frame selection (1 per second), the training set has 3,943,375 frames and the test set has 7,273,996 frames. Different image descriptors were evaluated for extracting features from such frames. The descriptors for encoding color properties are: Auto Color Correlogram (ACC) [18], Color Coherent Vector (CCV) [19], Border/Interior pixel Classification (BIC) [20], and Global Color Histogram (GCH) [21]. The texture descriptors evaluated are: Generic Fourier Descriptor (GFD) [22] and Haar-Wavelet Descriptor (HWD) [23]. For more details regarding those image descriptors, please refer to [24].

The experiments are divided into three phases. The first one evaluates the genre classifier. The second one evaluates the BoG representation for video genre retrieval and the last one evaluates the BoG representation for video event retrieval.

1) Evaluation of the genre classifier: The evaluation of the genre classifier is important because the quality of the final BoG vector depends on the quality of this classifier. If the genre classifier classifies the frames in wrong genres, the BoG vector will not reflect the correct distribution of video genres. It would be similar to have a bad coding step in traditional visual dictionaries of quantized local features: wrong visual word labels would be assigned to image patches, resulting in a bad bag of visual words. Therefore, the BoG model depends on a good genre classifier.

To create the visual dictionary, we trained a linear SVM ($c = 1.0$) using features extracted from the training videos. The genre (label) of each training frame is the same of the video from where it was extracted. The training of the genre classifier was based on randomly selecting the same number N of frames per genre. We varied N in 100, 500, and 800 frames per genre. The remaining frames were used for testing. It is worth mentioning the amount of frames used in this evaluation: almost 4 million of the training videos and more than 7.2 million of the test videos (no frames of the test videos were used for training the genre classifier). For running SVM, we used the LIBSVM package[3] [25].

Figure 2 presents the classification accuracy for the evaluated descriptors. Notice that the classification accuracies are low for all the descriptors, creating a very challenging scenario for the BoG model, as we explained previously. The best results were obtained for the SVM model learned on 800 training frames per class. This model was used for representing the test videos using the BoG approach in the following experiments.

2) Evaluation of the BoG representation: The following experiments evaluate the BoG model for video genre retrieval. Each video in the test set was represented by a bag of genres using the genre classifiers learned on the training step. With the BoG of each video, a given test video was used as query for the rest of the videos in the test set, which were ranked according to the Euclidean (L_2) distance between their BoGs. For each genre, around five percent of the test videos were randomly selected and used as queries. Five replications were performed in order to ensure statistically sound results. Presented results refer to the average scores and their respective 99% confidence intervals, which were computed based on the mean and standard deviation of each replication.

We compared the BoG approach against with two baselines: Histogram of Motion Patterns (HMP) [1] and Bag of Scenes (BoS) [10]. To make a fair comparison, these approaches were configured with their best settings based

[2]http://blip.tv (As of May 2016).

[3]http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (As of May 2016)

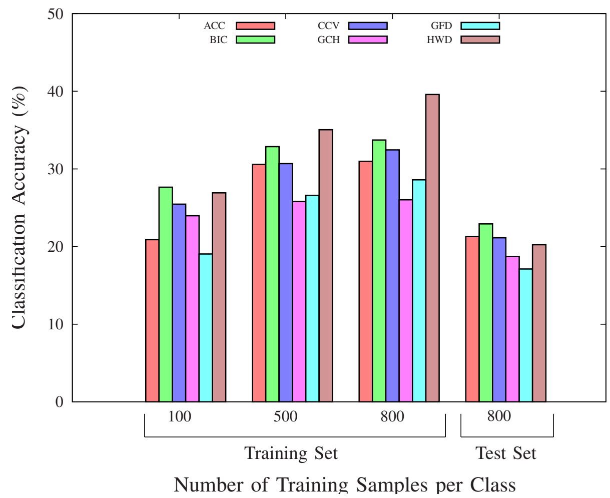


Fig. 2. Evaluation of the genre classifier. All descriptors generated low discriminating genre classifiers (accuracy below 50%), creating a challenging scenario for the BoG model.

on the results reported in [26]. The distance function used for feature comparison is the Euclidean (L_2) distance. The retrieval effectiveness was assessed using the precision at the top 10 retrieved items (P@10) and Mean Average Precision (MAP).

In Figure 3, we compare the BoG representations and the baseline methods with respect to the MAP and P@10 measures. As we can observe, the performance of the BoG representations are slightly better considering the MAP measure. MAP is a good indication of the effectiveness considering all positions of obtained ranked lists. P@10, in turn, focuses on the effectiveness of the methods considering only the first positions of the ranked lists.

The BoG approach achieved the best scores using BIC as the frame descriptor (used as basis for the genre classifier). Notice that BoG$_{BIC}$ performs better than the baseline methods for MAP, however the same does not happen for P@10. BIC was the best descriptor for the genre classifier in the test set (see Section IV-A1), making it also better for generating the BoG vector.

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 of each class 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.

Table I presents the 99% confidence intervals of the differences between BoG$_{BIC}$ (the best configuration of BoG) and the baseline methods for the MAP and P@10 measures, respectively. Notice that the confidence intervals for BoG$_{BIC}$ and BoS include zero and, hence, the differences between those approaches are not significant at that confidence level. On the other hand, the performance of BoG$_{BIC}$ and HMP are not statistically different for MAP, whereas BoG$_{BIC}$ performs worse than HMP for P@10. This method is based on motion information and, hence, it does not consider visual properties of video frames in an independent manner.

TABLE I

PAIRED t-TEST COMPARING THE BEST BOG CONFIGURATION AND THE BASELINES. WE CAN NOTE INTERVALS CROSSING THE ZERO FOR BOG$_{BIC}$ AND BOS, INDICATING NO STATISTICAL DIFFERENCE BETWEEN METHODS. FOR BOG$_{BIC}$ VERSUS HMP, HMP IS BETTER FOR P@10.

Approach	**MAP**		**P@10**	
	min.	**max.**	**min.**	**max.**
BoG$_{BIC}$ - BoS	-0.018	0.018	-0.063	0.014
BoG$_{BIC}$ - HMP	-0.074	0.007	-0.232	-0.079

Figure 4 compares the individual scores obtained for each class in terms of MAP and P@10 measures. It is interesting to note the differences in responsiveness of the different approaches with respect to each of the genres. For MAP, BoG$_{BIC}$ performs better than the baseline methods for most of the classes (13 out of 26). For P@10, BoG$_{BIC}$ provides a good discriminative power on genres like *"school and education"* and *"web development and sites"*.

The key advantage of the BoG model is its computational efficiency in terms of space occupation and similarity computation time. In our experiments, the BoG vector corresponds to a 26-bin histogram, which represents a reduction of 74% in relation to the BoS vector (100-bin histogram) and is two orders of magnitude smaller than the HMP vector (6075-bin histogram), making our approach more suitable for real-time processing.

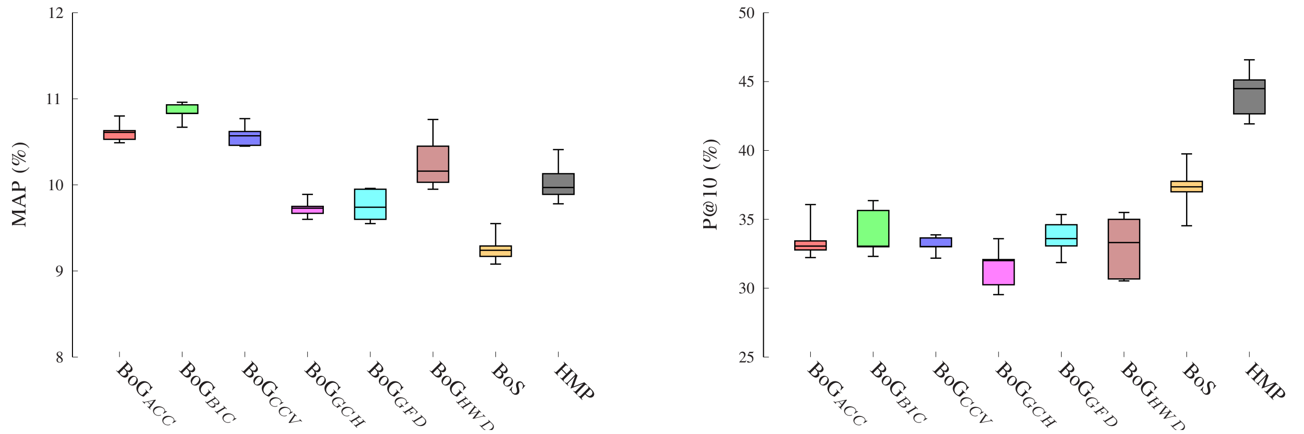
Although the effectiveness the BoG approach is not superior to the baseline methods, the obtained results show the potential of the idea. As we explained previously, the quality of the genre classifier is important for the BoG quality. Our genre classifiers obtained less than 50% of accuracy in the training set and less than 30% in the test set, probably limiting the quality of the BoG representation. Another limitation is the dataset used. As all the frames of a video have the same label, visually different frames may be of the same genre, harming the classifier.

B. Video Event Retrieval

Also, we carried out this study on the EVVE (EVent VidEo) dataset[4]: an event retrieval benchmark introduced by Revaud et al. [27]. The dataset is composed of 2,995 videos (166 hours) collected from YouTube[5]. 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 II. It is a challenging benchmark since the events are localized in both time and space, for instance, the event 1 refers to the great riots and strikes that happened in the end

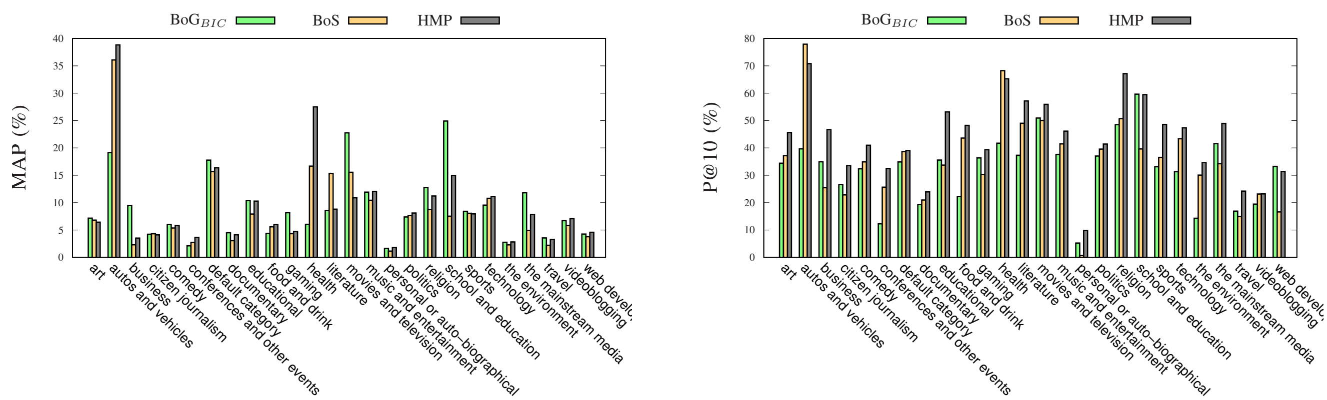
[4]http://pascal.inrialpes.fr/data/evve/ (As of May 2016).

[5]http://www.youtube.com (As of May 2016).



(a) MAP (b) P@10

Fig. 3. Results for video genre retrieval comparing BoG with the baselines in terms of MAP and P@10. BoG$_{BIC}$ obtained the best MAP score.



(a) MAP (b) P@10

Fig. 4. MAP and P@10 scores obtained for each genre.

TABLE II

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**

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.

Our experiments followed the official experimental protocol created by [27]. Initially, each video in the dataset was represented by a BoG. With the BoG of each video, a given query video was used to retrieve the rest database videos, which were ranked according to the Euclidean (L2) distance between their BoGs. Finally, we used the dataset official tool to evaluate the retrieval results[6].

In this experiment, we used BoG_{BIC} to represent videos, which was the approach that achieved the best scores for video genre retrieval. Our intend here is to verify if the BoG representation can perform well in a different scenario.

We compared the BoG_{BIC} approach against three baselines [27]: Mean-MultiVLAD (MMV), CTE (Circulant Temporal Encoding) and a combination of both methods, known as MMV+CTE. Also, we considered the variations of MMV with the following hyper-pooling functions [28]: k-means, partial k-means (PKM), sign of stable componentes (SSC), KD-Tree and Fisher Vectors. To make a fair comparison, these approaches were selected with their best performance based on the results reported in [27], [28].

In Figure 5, we compare the BoG_{BIC} representation and the baseline methods with respect to the avg-mAP. As we can observe, the performance of the BoG_{BIC} representation outperformed all baseline methods by a large margin.

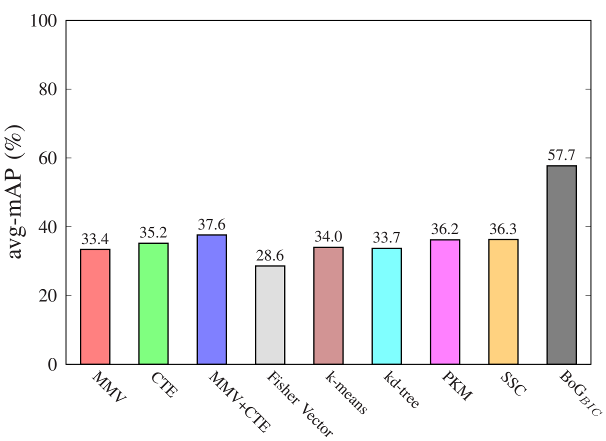


Fig. 5. Performance of different methods for event retrieval on EVVE dataset.

The results were also compared by event, as shown in Table III. One can notice that BoG_{BIC} representation performed better than the baseline methods for most of the events (11 out of 13). For some events, the difference in favor of our method is very large, like in events 4, 5, 8, and 12.

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.

Table IV presents the 95% confidence intervals of the differences between BoG_{BIC} and the baseline methods for the mAP measures. Notice that the confidence intervals for BoG_{BIC} and the baseline methods are always positive, indicating that BoG_{BIC} outperformed those approaches.

[6]http://pascal.inrialpes.fr/data/evve/eval_evve.py (As of May 2016).

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

Event ID	MMV	CTE	MMV+CTE	BoG$_{BIC}$
1	23.90	13.90	24.60	**25.32**
2	19.90	16.60	20.20	**42.58**
3	8.70	12.80	11.10	**45.51**
4	12.60	10.80	13.20	**79.41**
5	23.40	26.20	26.00	**47.20**
6	33.80	41.30	39.40	**56.66**
7	12.40	25.20	21.20	**33.63**
8	25.40	25.70	28.10	**74.71**
9	53.10	**80.30**	69.40	28.05
10	45.50	40.90	**48.60**	45.11
11	77.30	71.40	77.40	**89.04**
12	36.60	29.70	37.10	**98.57**
13	60.40	69.30	71.90	**84.87**
avg-mAP	33.40	35.20	37.60	**57.74**

TABLE IV
PAIRED T-TEST COMPARING BoG_{BIC} AND THE BASELINES. AS INTERVALS ARE ABOVE ZERO, WE CAN SAY THAT BoG_{BIC} OUTPERFORMED THE BASELINES WITH STATISTICAL SIGNIFICANCE.

Approach	mAP	
	min.	max.
BoG$_{BIC}$ - MMV	0.091	0.398
BoG$_{BIC}$ - CTE	0.034	0.407
BoG$_{BIC}$ - MMV+CTE	0.031	0.373

According to the analysis of BoG_{BIC} results per event, one of the worst results happened on the event 1. On the other side, the best event was obtained on the event 13. We made a visual analysis at the videos to help to understand the differences.

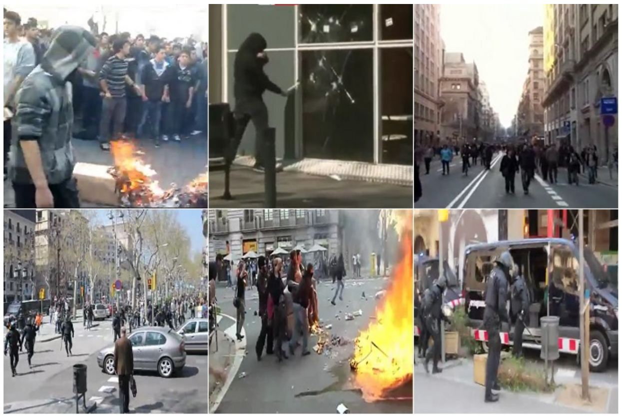
In case of the event 1 (see Figure 6), it is possible to see lots of riots and strikes at different places and moments. There are scenes showing police, fire, cars, and crowd in almost all the videos (Figure 6(b)). Thus, it is difficult to identify only videos of the austerity riots that occurred in Barcelona at the end of March, 2012 (Figure 6(a)). As shown in Table III, all the methods performed below 25% for this event.

But, in case of the event 13 (see Figure 7), there are lots of similar positive videos, specially recorded at the entrance of the ride, as shown in Figure 7(a). This scene is repeated in many videos and probably helped our method. Negative videos do not contain the same entrance, as shown in Figure 7(b).

We believe that our method outperformed the baseline methods because the proposed BoG representation carries semantic information. But, on the other side, our method does not include temporal information and we think such feature is important to recognize some types of events.

V. CONCLUSIONS

In this paper, we presented a new video representation for video retrieval, named as Bag of Genres. This representation model relies on a dictionary of genres, which is created from a genre classification model learned on the training frames. Different from traditional dictionaries based on local features



(a) Positive Videos

(b) Negative Videos

Fig. 6. Examples of video frames from Event 1 (Austerity riots in Barcelona, 2012), which was one of the events that BoG performed worst.

(a) Positive Videos

(b) Negative Videos

Fig. 7. Examples of video frames from Event 13 (Jurassic Park ride in Universal Studios theme park), which was one of the events that BoG performed best.

(e.g., SIFT or STIP), here, visual words correspond genre-labeled regions of the classification space. Therefore, each dimension of the feature space spanned by such a model is associated to a semantic concept.

Our approach was validated in the dataset of MediaEval Tagging Task of 2012. Our experiments evaluated the importance of the genre classifier in the model as well as the quality of the BoG representation. In these experiments, the BoG model has performed well despite the low accuracy of the genre classifier. The results demonstrated that our approach performs similar to state-of-the-art methods, but using a much more compact representation. Also, we tested the best configuration of the BoG model to retrieve videos by event on the EVVE dataset. The results show that our approach outperformed state-of-the-art solutions.

We can think about ways of improving the BoG model. For instance, a smarter strategy for feature extraction and classification may enable to create more informative visual dictionaries and, hence, improve the video representation.

Future work includes the evaluation of other methods for feature extraction, as well as perform an extensive study on classification strategies to be used in the creation of visual dictionaries. We also would like to evaluate the use of a dataset of scene images to create the genre classifier.

Acknowledgment

The authors would like to thank CAPES, CNPq, and FAPESP (grant #2016/06441-7) for funding.

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] 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 *Annual Conference on Neural Information Processing Systems (NIPS'10)*, 2010, pp. 1378–1386.

[3] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *International Conference on Computer Vision*, 2009, pp. 1365–1372.

[4] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[5] 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.

[6] S. Albaradei, Y. Wang, L. Cao, and L.-J. Li, “Learning mid-level features from object hierarchy for image classification,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2014, pp. 235–240.

[7] S. Bhattacharya, M. Kalayeh, R. Sukthankar, and M. Shah, “Recognition of complex events: Exploiting temporal dynamics between underlying concepts,” in *Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2243–2250.

[8] 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. 1-2, pp. 59–81, 2014.

[9] 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’12)*, 2012, pp. 1234–1241.

[10] 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.

[11] 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.

[12] 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.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” *CoRR*, vol. abs/1312.6229, 2013. [Online]. Available: http://arxiv.org/abs/1312.6229

[14] 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.

[15] S. Schmiedeke, C. Kofler, and I. Ferrané, “Overview of mediaeval 2012 genre tagging task,” in *Working Notes Proceedings of the MediaEval 2012 Workshop (MEDIAEVAL’12)*, 2012.

[16] B. E. Ionescu, K. Seyerlehner, I. Mironica, C. Vertan, and P. Lambert, “An audio-visual approach to web video categorization,” *Multimedia Tools Applications*, vol. 70, no. 2, pp. 1007–1032, 2014.

[17] I. Mironica, B. Ionescu, P. Knees, and P. Lambert, “An in-depth evaluation of multimodal video genre categorization,” in *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI’13)*, 2013, pp. 11–16.

[18] J. Huang, R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, “Image indexing using color correlograms,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR’97)*, 1997, pp. 762–768.

[19] G. Pass, R. Zabih, and J. Miller, “Comparing images using color coherence vectors,” in *ACM International Conference on Multimedia (ACM-MM’96)*, 1996, pp. 65–73.

[20] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, “A compact and efficient image retrieval approach based on border/interior pixel classification,” in *ACM International Conference on Information and Knowledge Management (CIKM’02)*, 2002, pp. 102–109.

[21] M. J. Swain and B. H. Ballard, “Color indexing,” *Int. J. Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[22] D. Zhang and G. Lu, “Shape-based image retrieval using generic fourier descriptor,” *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 825–848, 2002.

[23] C. E. Jacobs, A. Finkelstein, and D. Salesin, “Fast multiresolution image querying,” in *ACM International Conference Computer Graphics and Interactive Techniques (SIGGRAPH’95)*, 1995, pp. 277–286.

[24] O. A. B. Penatti, E. Valle, and R. S. Torres, “Comparative study of global color and texture descriptors for web image retrieval,” *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 359–380, 2012.

[25] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[26] J. Almeida, D. C. G. Pedronette, and O. A. B. Penatti, “Unsupervised manifold learning for video genre retrieval,” in *Iberoamerican Congress on Pattern Recognition (CIARP’14)*, 2014, pp. 604–612.

[27] J. Revaud, M. Douze, C. Schmid, and H. Jégou, “Event retrieval in large video collections with circulant temporal encoding,” 2013, pp. 2459–2466.

[28] M. Douze, J. Revaud, C. Schmid, and H. Jégou, “Stable hyper-pooling and query expansion for event detection,” 2013, pp. 1825–1832.