Unsupervised Manifold Learning for Video Genre Retrieval

Jurandy Almeida[1], Daniel C. G. Pedronette[2], and Otávio A. B. Penatti[3]⋆

[1] Institute of Science and Technology, Federal University of São Paulo – UNIFESP
12231-280, São José dos Campos, SP – Brazil
jurandy.almeida@unifesp.br

[2] Dept. of Statistics, Applied Mathematics and Computation, São Paulo State University – UNESP
13506-900, Rio Claro, SP – Brazil
daniel@rc.unesp.br

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

Abstract. This paper investigates the perspective of exploiting pairwise similarities to improve the performance of visual features for video genre retrieval. We employ manifold learning based on the reciprocal neighborhood and on the authority of ranked lists to improve the retrieval of videos considering their genre. A comparative analysis of different visual features is conducted and discussed. We experimentally show in the dataset of 14,838 videos from the MediaEval benchmark that we can achieve considerable improvements in results. In addition, we also evaluate how the late fusion of different visual features using the same manifold learning scheme can improve the retrieval results.

Keywords: video genre retrieval; ranking methods; manifold learning

1 Introduction

Recent advances in technology have increased the availability of video data. This has spurred great interest in efficient systems for managing video material. The main challenge of those systems is to identify and select only relevant information according to user needs.

In the last years, visual content-based systems have emerged as an alternative to overcome the limitations of traditional text-based systems. They rely on extracting low-level features from videos and determining similarity between them by computing distances between feature vectors.

In spite of all the advances, the "semantic gap"[1] is still an open problem. To bridge this gap, various unsupervised strategies [6, 7, 12] can be employed. Once videos (and other multimedia objects) live in a much lower-dimensional intrinsic space than the feature vectors that represent them, capturing and exploiting the intrinsic manifold structure therefore becomes a central problem in the vision community [4]. Manifold learning approaches have been successfully used in several scenarios [5–7]. Such approaches aim at computing new distances between

⋆ Thanks to Brazilian agencies FAPESP (grant 2013/08645-0), CNPq, and CAPES for funding. We also thank the RECOD Lab from Institute of Computing at University of Campinas for the infrastructure.

[1] Videos with high feature similarities may be different in terms of user perception.

objects, exploring the dataset manifold [13] and the reciprocal neighborhood of ranked lists.

The main purpose of this paper is to show improvements of genre-based video retrieval using a manifold learning approach. We have used recently proposed manifold learning approaches [6] in the dataset of MediaEval Genre Tagging Task of 2012 [10]. The application of such approaches using a set of several visual features extracted from the videos can improve the results considerably.

The remainder of this paper is organized as follows. Section 2 describes the visual features used to represent the videos. Section 3 presents the unsupervised manifold learning approaches. Section 4 shows the experiments and results and Section 5 concludes the paper.

2 Visual Features

To encode video visual properties, we have used three main approaches. Two of them are based on video frames and do not consider transitions between them: *bag of visual words* and *bag of scenes* [8]. The other approach specifically encodes motion information by using *histogram of motion patterns* [1].

2.1 Bag of Visual Words (BoVW)

Nowadays, bags of visual words are very popular in the computer vision literature [3]. They represent visual content by statistical information of local patterns, encoding the occurrences of quantized local features. Local features, like SIFT and SURF, tend to be very specific, therefore quantizing their feature space increases the generality of descriptions making BoVW representations appropriate for a variety of applications. The feature-space quantization creates the so-called visual dictionary or visual codebook. The most important steps for computing the BoVW representation after the dictionary creation are *coding* [11] and *pooling* [3]. *Hard* and *soft* assignment are usually employed for coding, while *average* and *max* are common operations for pooling features in the final BoVW vector.

In this paper, we extracted bags of visual words from videos by performing pooling in two stages. Initially, we considered frames isolated and computed the BoVW vector for each frame. Then, we performed another pooling operation over the BoVW of all the frames of a given video. This second pooling operation generated the BoVW for the video. To differentiate from the other features in this paper, we called it as *pooling over pooling* (PoP). $\text{PoP}_{1000}^{soft1+avg,max}$ means that a codebook of 1000 visual words was used for the local features, soft assignment ($\sigma = 1$) and average pooling were used to compute the BoVW of each frame, and max pooling was used to combine frame BoVWs.

2.2 Bag of Scenes (BoS)

Bag of scenes is an approach for encoding video visual properties [8]. It is based on a dictionary of scenes which is composed of a set of scenes of interest. Such dictionary can be created similarly to dictionaries based on local features (e.g., SIFT). An important advantage of the bag-of-scenes model is that the dictionary is composed of visual words carrying more semantic information than the traditional dictionaries based on local descriptions. In the dictionary of scenes, each visual word can be more clearly associated with a visual concept than a local patch. As a consequence, the bag-of-scenes feature space has one dimension for each semantic concept, making it easier to detect the presence or absence of the concept in the video feature vector.

For creating the bag of scenes, several coding and pooling strategies used in the BoVW representation can be used, like *hard* and *soft* assignment [11], *average* and *max* pooling [3], for instance. In this paper, we use multiple configurations for computing the BoS of a video, like varying the dictionary size, using hard or soft assignment and average or max pooling. To differentiate them, we use the following abbreviation $\text{BoS}_{1000}^{soft2+avg}$, which says that soft assignment ($\sigma = 2$) and average pooling were used to compute the bag of scenes of a video, using a dictionary of 1000 scenes.

2.3 Histogram of Motion Patterns (HMP)

Besides encoding visual properties using visual dictionaries, we also adopted a simple and fast algorithm to compare videos [1]. It consists of three main steps: (1) partial decoding; (2) feature extraction; and (3) signature generation.

For each frame of an input video, motion features are extracted from the video stream. For that, 2 × 2 ordinal matrices are obtained by ranking the intensity values of the four luminance (Y) blocks of each macro block. This strategy is employed for computing both the spatial feature of the 4-blocks of a macro block and the temporal feature of corresponding blocks in three frames (previous, current, and next). Each possible combination of the ordinal measures is treated as an individual pattern of 16-bits (i.e., 2-bits for each element of the ordinal matrices). Finally, the spatio-temporal pattern of all the macro blocks of the video sequence are accumulated to form a normalized histogram.

3 Unsupervised Manifold Learning for Video Retrieval

The effectiveness of multimedia retrieval applications depends on different steps of the retrieval process. Besides the visual features used, the distance measure adopted plays an important role, directly affecting the quality of retrieved results. In general, multimedia retrieval systems consider only pairwise analysis, that is, compute similarity/distance measures considering only pairs of objects. Since only pairwise distances are considered, information about the neighborhood of the query is ignored.

On the other hand, the user perception considers the query specification and responses in a given *context*. Therefore, an effective distance measure should consider the similarity among the query and retrieved objects in the context of the whole collection [13, 6]. In view of that, several approaches have been proposed [5, 13, 7, 4] aiming at replacing pairwise similarities by more global affinity measures.

Manifold learning approaches can be used for learning global affinity measures. The main motivation of manifold learning consists in computing new distances between objects that correspond to geodesic distances on the dataset manifold [13]. The new distances are estimated considering a walk along the geometric structure of the dataset. In this paper, we use a recent proposed unsupervised manifold learning approach [6] based on Reciprocal kNN Graphs, described in next section. The main motivation consists in using the manifold learning method for computing a new and more accurate distance among videos, improving the effectiveness of video retrieval tasks.

3.1 Unsupervised Manifold Learning by Reciprocal kNN Graphs

The Unsupervised Manifold Learning by Reciprocal kNN Graphs [6] is based on the information given by top-k positions of the ranked lists, which encode relevant contextual information. Given a query video, the ranked lists define relationships not only between pairs of videos (as distance functions), but also among all the videos in the ranked list. The manifold learning algorithm [6] analyzes the dataset structure by considering the reciprocal references among objects at top positions of their ranked lists.

The *Reciprocal kNN Graph* method exploits the contextual information in ranked lists using three main strategies [6]:

- **Reciprocal Neighborhood:** the k-reciprocal nearest neighborhood, is a much stronger indicator of similarity than the unidirectional nearest neighborhood [9], reducing the risk of false positives at top positions of ranked lists.
- **Collaborative Ranking:** aiming at computing a more global affinity, the method employs a collaborative analysis. The motivation consists in the fact that a ranked list can provide useful information for improving effectiveness of other ranked lists [7].
- **Authority of Ranked Lists:** the manifold learning approach computes a score for measuring the graph's density that represents the reciprocal references among objects at top positions of the ranked list. The score is used to estimate the authority of a given ranked list for collaborating with other ranked lists.

At each iteration, the manifold learning method computes a new distance among objects, considering the reciprocal neighborhood analysis. Based on new distances, new ranked lists are computed repeating the process until convergence. The method can also be used for distance fusion, aiming at combining distances computed by different visual features.

4 Experimental Evaluation

This section presents the experimental evaluation conducted and discusses the obtained results. Section 4.1 and 4.2 describe, respectively, the dataset and effectiveness measures used. Section 4.4 discusses the impact of parameters. Section 4.3 presents the effectiveness results for all visual features and Section 4.5 discusses the results in combination tasks.

4.1 Dataset

In this work, we use a benchmarking dataset provided by the MediaEval 2012 organizers for the Genre Tagging Task [10]. The dataset is composed of 14,838 videos (3,288 hours) collected from the blip.tv[2]. 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

[2] `http://blip.tv` (as of May, 2014).

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 an (116) and default category (2349, comprises videos that cannot be assigned to any of the previous categories). The main challenge of this scenario is the high diversity of genres, as well as the high variety of visual contents within each genre category.

4.2 Effectiveness Measures

The effectiveness of each approach was assessed using the metrics of Precision and Recall [2]. Precision is the ratio of relevant videos in the retrieved set of videos. Recall is the ratio of relevant videos retrieved in relation to the total number of relevant videos in the database. There is a trade-off between Precision and Recall, i.e., increasing Recall may decrease Precision and vice versa. For this reason, we consider unique-value measurements in the validation: Mean Average Precision (MAP), which is the mean of the precision scores obtained at the ranks of each relevant video; and Precision at 10 (P@10), which is the average precision after 10 videos are returned. 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.

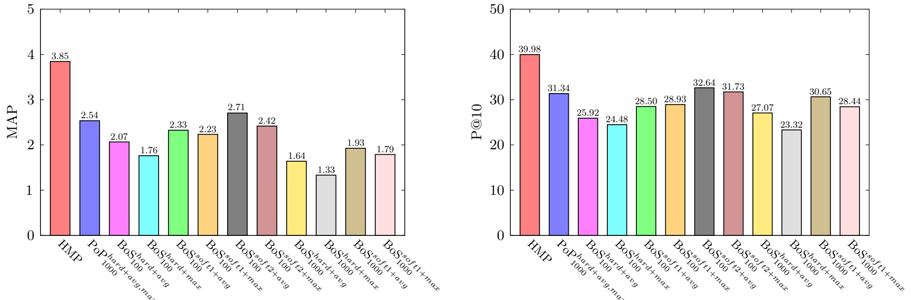
It is also important to mention the distance functions used to generate the original rankings for each visual feature: Euclidean distance for PoP and BoS and histogram intersection for HMP.

All experiments were conducted considering all the videos from the dataset as queries. Results reported (for both MAP and P@10 measures) represent the average of all the videos. Since our objective is to analyze the gains obtained by the use of the manifold learning algorithm [6], we report the relative gain, which is given by the absolute gain divided by the initial effectiveness score.

4.3 General Effectiveness Results

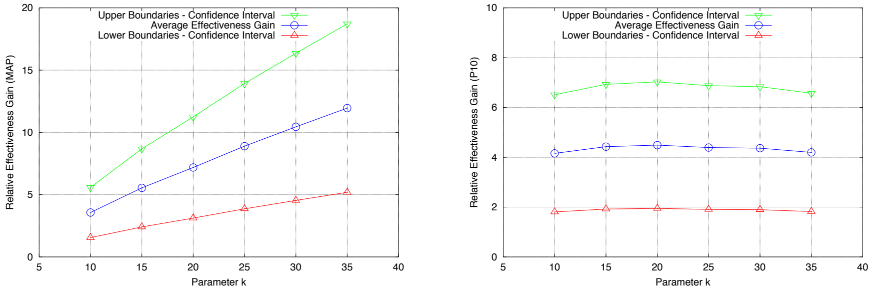
In this section, we aim at presenting and discussing the overall effectiveness results obtained by the use of the manifold learning algorithm [6] considering all the 12 visual features. In Figure 1, we compare the visual features with respect to the MAP and P@10 measures, respectively. Notice that, the initial effectiveness scores are low for both measures (MAP and P@10), which turns more challenging the use of unsupervised approaches.

Figure 2 presents the average relative gains for all the evaluated descriptors and different values of k, considering the MAP and P@10 measures (blue lines). The 95% confidence intervals are also reported, in green and red lines for upper and lower boundaries, respectively. We can observe very significant average gains for MAP, reaching +11.95%. For P@10, the average gains are lower, but still significant, reaching +4.49%.



(a) MAP (b) P@10

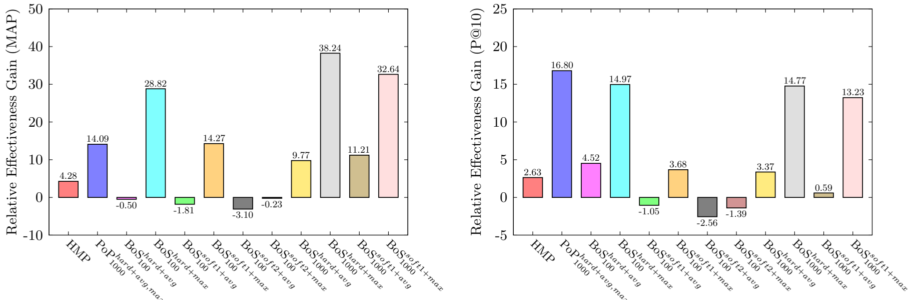
Fig. 1. MAP and P@10 scores obtained by each of the visual features.



(a) MAP (b) P@10

Fig. 2. Average effectiveness gains considering all the visual features.

Figure 3 presents the individual results obtained by the visual features for the best choice of the parameter k considering the MAP and P@10 measures, respectively. Notice the relative gains obtained for both measures, reaching +38.24% for MAP and +16.80% for P@10. The significant gains obtained demonstrates the usefulness of manifold learning even in this challenging scenario.



(a) MAP (b) P@10

Fig. 3. The best results obtained for each of the visual features.

4.4 Analysis of Parameters

The manifold learning algorithm [6] requires two parameters: *(i)* k: number of neighbors considered in the unsupervised learning process; and *(ii)* ϵ: a convergence threshold parameter. We used the same threshold parameter value ($\epsilon = 0.0125$) used in [6]. In this section, we aim at evaluating the impact of the parameter k on the video retrieval effectiveness. We evaluated the effectiveness scores and the relative gain for different values of k. For that, we considered only the HMP [1] descriptor as it presented the highest MAP and P@10 scores before employing manifold learning.

In Figure 4, we show the relative gain of the MAP and P@10 measures using the manifold learning as the parameter k increases. We can see that, as more elements are analyzed in each ranked list (larger k), more improvement is obtained, until reach a peak. This is an expected behavior, because increasing k consists in analyzing a larger reciprocal neighborhood, which aggregates more information. From a certain k, however, non-relevant results are considered and the gain decreases.

We can also see that, although the best k is different depending on the measure (MAP or P@10), the value is small for both measures. As MAP is related to the quality of results when retrieving all the relevant videos from the dataset, it benefits of learning from a larger k. On the other hand, as P@10 refers to the quality of only the top results, a small set is enough.

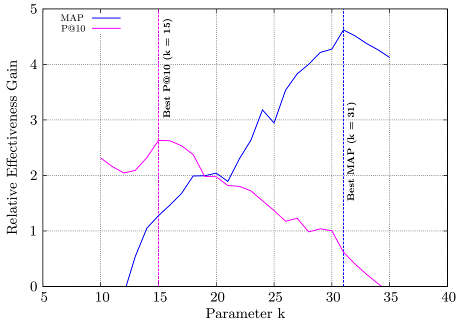


Fig. 4. Effectiveness gains obtained for the HMP by varying the parameter k.

4.5 Combination of Visual Features

The manifold learning algorithm [6] was also evaluated for distance fusion. For that, we considered the visual features which have presented the best effectiveness scores (HMP [1] and PoP [8]). We also considered the value of k ($k = 31$) which presented the best results for the HMP [1] descriptor.

Table 1. Effectiveness results of manifold learning approach for feature combination.

Algorithm	**Initial MAP**	**Manifold Learning**	**Relative Gain**	**Initial P@10**	**Manifold Learning**	**Relative Gain**
HMP [1]	3.85%	4.02%	+4.42%	39.98%	40.22%	+0.60%
PoP [8]	2.54%	2.87%	+12.99%	31.34%	36.55%	+16.62%
HMP [1]+PoP [8]	-	**4.33%**	+12.47%	-	**45.24%**	+13.16%

Table 1 presents the results for the combination and for each descriptor isolated. The relative gain of the combination was computed over the best visual feature (before manifold learning). We can observe that the combination achieved the best effectiveness scores, for both MAP and P@10, reaching a relative gain of +13.16% over the best feature.

5 Conclusions

This paper presented an evaluation of manifold learning for video genre retrieval. We employed manifold learning approaches over a set of visual features extracted from the video dataset used in the MediaEval Genre Tagging Task of 2012, which contains more than 14 thousand videos. This dataset is very challenging for visual descriptors and the results obtained before manifold learning are quite low. Even in this scenario of low precision values, the manifold learning approaches could largely improve the results, in some cases having an accuracy gain of more than 35%. We also notice the successful use of manifold learning for distance fusion, reaching gains of more than 13% over the best isolated feature in some cases. Those results indicate the importance of considering the dataset structure for reducing the semantic gap and for improving video retrieval. As future work, we intend to evaluate the combination of other features.

References

1. Almeida, J., Leite, N.J., Torres, R.S.: Comparison of video sequences with histograms of motion patterns. In: ICIP. pp. 3673–3676 (2011)
2. Bimbo, A.: Visual information retrieval. Morgan Kaufmann Publishers Inc. (1999)
3. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. pp. 2559–2566 (2010)
4. Jiang, J., Wang, B., Tu, Z.: Unsupervised metric learning by self-smoothing operator. In: ICCV. pp. 794–801 (2011)
5. Pedronette, D.C.G., Almeida, J., Torres, R.S.: A scalable re-ranking method for content-based image retrieval. Information Sciences 265, 91–104 (2014)
6. Pedronette, D.C.G., Penatti, O.A.B., da S. Torres, R.: Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. Image and Vision Computing 32(2), 120 – 130 (2014)
7. Pedronette, D.C.G., da S. Torres, R.: Image re-ranking and rank aggregation based on similarity of ranked lists. Pattern Recognition 46(8), 2350–2360 (2013)
8. Penatti, O.A.B., Li, L.T., Almeida, J., da S. Torres, R.: A visual approach for video geocoding using bag-of-scenes. In: ICMR. pp. 1–8 (2012)
9. Qin, D., Gammeter, S., Bossard, L., Quack, T., van Gool, L.: Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: CVPR. pp. 777 –784 (2011)
10. Schmiedeke, S., Kofler, C., Ferrané, I.: Overview of mediaeval 2012 genre tagging task. In: MediaEval (2012)
11. van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. TPAMI 32(7), 1271–1283 (2010)
12. Yang, X., Prasad, L., Latecki, L.: Affinity learning with diffusion on tensor product graph. TPAMI 35(1), 28–38 (2013)
13. Yang, X., Latecki, L.J.: Affinity learning on a tensor product graph with applications to shape and image retrieval. In: CVPR. pp. 2369–2376 (2011)