Scene-driven Retrieval in Edited Videos using Aesthetic and Semantic Deep Features

Lorenzo Baraldi, Costantino Grana and Rita Cucchiara
Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia
Via P. Vivarelli, 10, Modena MO 41125, Italy
name.surname@unimore.it

ABSTRACT

This paper presents a novel retrieval pipeline for video collections, which aims to retrieve the most significant parts of an edited video for a given query, and represent them with thumbnails which are at the same time semantically meaningful and aesthetically remarkable. Videos are first segmented into coherent and story-telling scenes, then a retrieval algorithm based on deep learning is proposed to retrieve the most significant scenes for a textual query. A ranking strategy based on deep features is finally used to tackle the problem of visualizing the best thumbnail. Qualitative and quantitative experiments are conducted on a collection of edited videos to demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis and indexing

Keywords

Video retrieval, Thumbnail selection, Ranking

1. INTRODUCTION

Suppose to search for a given content in a large video collection, which contains long edited videos with different subjects and heterogeneous content, like a collection of documentaries or movies. In this context, users would like to have a quick overview of results, even with a low precision, but capable to give a glance of what can be associated with a query for a further manual refining. Examples are in advertisement where video are re-used to find interesting sequences, in education and edutainment to enrich textual explanations with visual suggestions, in magazine editing, in broadcast-to-web presentations, and also in web search engines.

Nowadays, retrieval is changing towards a greater focus on *aesthetic quality*, a subjective aspect difficult to quantify.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR 2016 New York, USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

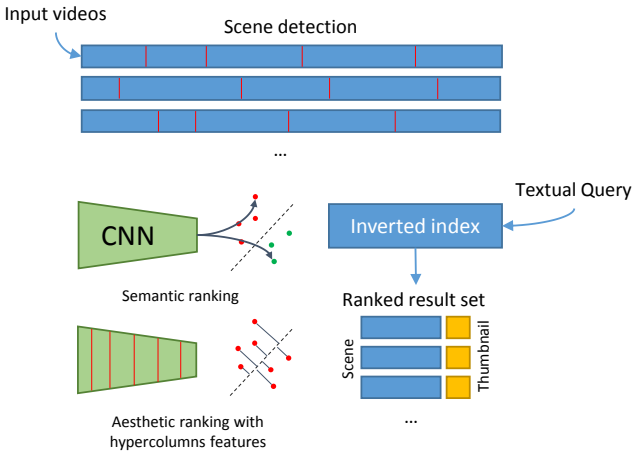


Figure 1: Overview of the proposed method. Given a collection of videos and a textual query, we retrieve a ranked list of the most significant parts (*scenes*) according to both semantics and aesthetic quality. Each retrieved scene is presented with an appropriate thumbnail.

Datta *et al.* [5] assessed that modeling aesthetics of images is an important open problem, and it is still not solved. It concerns in general with the kind of emotions a picture arises in people, or more simply in beauty-related of images or videos.

This is an old story: Plato, in *Symposium*, discusses his idea of beauty, that comes from the perception of objects, their proportions, their harmony or unity among the parts, in the evenness of the line and purity of color. This Greek ideal permeates most of the occidental concepts of beauty and the current aesthetic theories, and affects as well theories on user interface designs and, recently, on retrieval too. Google, for instance, spent a large effort in changing the image search interface and the ranking, in order to convey not only the most relevant, abut also the most beautiful or fine results. Hongyi Li, associate product manager at Google, said that Google Images has been redesigned to provide "a better search experience, faster, more beautiful and more reliable"[1]. If image retrieval results are generally not only concerning the query but also ranked to have the more aesthetically valuable, this can be done also in video retrieval, where the complexity is higher. Moreover, also the granu-

[1]`https://googlewebmastercentral.blogspot.co.uk/2013/01/faster-image-search.html`

larity level could be changed: it is often the case, indeed, that long videos contain different parts and topics, hence an effective retrieval strategy should be able to recover the exact portion of the video the user is looking for.

In this paper we address the problem to provide both semantically and aesthetically valuable results of a query-by-text-retrieval in collections of long and heterogeneous video. Results are presented by thumbnails which recall the content of a video part associated with the specific search concept. Our proposal addresses three common drawbacks of the existing video retrieval pipelines. First, we do not rely on manually provided annotations, like descriptions or tags, which are expensive and not always available, and exploit solely the visual and audio content of the video. Secondly, we let the user search inside a video with a finer granularity level. Lastly, once a set of candidate results has been collected, each should be presented to the user with a thumbnail which is coherent with the query and aesthetically pleasant. To lower the granularity level of searches, we temporally segment each video into a set of semantically meaningful sequences. This task, which is known in literature as *scene detection*, results in a collection of scenes which have the property to be almost constant from a semantic point of view, and therefore constitute the ideal unit for video retrieval.

2. RELATED WORK

The process of producing thumbnails to represent video content has been widely studied. Most conventional methods for video thumbnail selection have focused on learning visual representativeness purely from visual content [10, 18]; however, more recent researches have focused on choosing query-dependent thumbnails to supply specific thumbnails for different queries. Craggs *et al.* [4] introduced the concept that thumbnails are surrogates for videos, as they take the place of a video in search results. Therefore, they may not accurately represent the content of the video, and create an *intention gap*, i.e. a discrepancy between the information sought by the user and the actual content of the video. To reduce the intention gap, they propose a new kind of animated preview, constructed of frames taken from a full video, and a crowdsourced tagging process which enables the matching between query terms and videos. Their system, while going in the right direction, suffers from the need of manual annotations, which are expensive and difficult to obtain.

In [13], instead, authors proposed a method to enforce the representativeness of a selected thumbnail given a user query, by using a reinforcement algorithm to rank frames in each video and a relevance model to calculate the similarity between the video frames and the query keywords. Recently, Liu *et al.* [14] trained a deep visual-semantic embedding to retrieve query-dependent video thumbnails. Their method employs a deeply-learned model to directly compute the similarity between a query and video thumbnails, by mapping them into a common latent semantic space.

On a different note, lot of work has also been proposed for video retrieval: with the explosive growth of online videos, this has become a hot topic in computer vision. In their seminal work, Sivic *et al.* proposed *Video Google* [21], a system that retrieves videos from a database via bag-of-words matching. Lew *et al.* [12] reviewed earlier efforts in video retrieval, which mostly relied on feature-based relevance feedback or similar methods.

Recently, concept-based methods have emerged as a popular approach to video retrieval. Snoek *et al.* [22] proposed a method based on a set of concept detectors, with the aim to bridge the semantic gap between visual features and high level concepts. In [2], authors proposed a video retrieval approach based on tag propagation: given an input video with user-defined tags, Flickr, Google Images and Bing are mined to collect images with similar tags: these are used to label each temporal segment of the video, so that the method increases the number of tags originally proposed by the users, and localizes them temporally. Our method, in contrast, does not need any kind of manual annotation, but is applicable to edited video only.

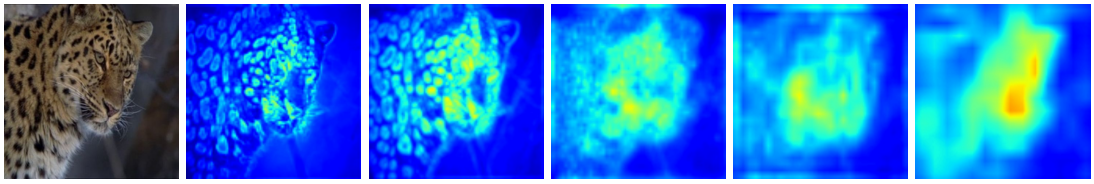
3. VISUAL-SEMANTIC RETRIEVAL

Given a set of videos $\mathcal{V}$, each decomposed into a set of scenes, and a query term q, we aim at building a function $r(q)$ which returns an ordered set of (video, scene, thumbnail) triplets. The retrieved scene must belong to the retrieved video, and should be as consistent as possible with the given query. Moreover, the returned thumbnail must belong to the given scene, and should be representative of the query as well as aesthetically remarkable.

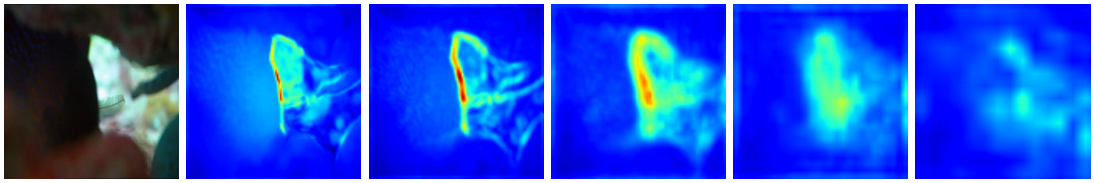
To detect whether a (video, scene, thumbnail) triplet should correspond to a query, we exploit visually confirmed concepts found in the transcript, as well as a measure of aesthetic quality. We parse the transcript of a video to identify candidate concepts, like objects, animal or people. Then, for each concept a visual classifier is created *on-the-fly* to confirm its presence inside the video, by means of an external corpus of images. Notice that when the transcript of video is not given, it can be easily replaced with the output of a standard speech-to-text software.

Scene detection To segment an input video into a set of coherent scenes, we apply the state-of-the-art algorithm described in [3]. Given a ground-truth temporal segmentation of a set of videos, [3] first runs a shot detector, then trains a Siamese Deep network to predict whether two shots should belong to the same temporal segment. Each branch of the Siamese network is composed by a Convolutional Neural Network (CNN) which follows the AlexNet architecture [11], and whose penultimate layer is concatenated with features extracted from the transcript of the video. The overall network is trained using a contrastive loss function, which computes the distance between two input shots. In test phase, distances between shots provided by the Siamese network are arranged into a similarity matrix, wich is then used together with Spectral Clustering to obtain the final scene boundaries.

Semantic concept detection Sentences in the corpus are parsed and words annotated as *noun*, *proper noun* and *foreign word* are collected with the Stanford CoreNLP part of speech tagger [6]. Each term is converted into its *lemmatized* form, so that nouns in singular and plural form are grouped together. Due to the huge variety of concepts which can be found in the video collection, the video corpus itself may not be sufficient to train detectors for the visual concepts. Therefore, we mine images from the Imagenet database [7], which contains images from more than 40.000 categories from the WordNet [17] hierarchy. Notice that our method, in principle, is applicable to any visual corpus, provided that it contains a sufficient large number



(a) Input image (b) `conv1*` (c) `conv2*` (d) `conv3*` (e) `conv4*` (f) `conv5*`



(g) Input image (h) `conv1*` (i) `conv2*` (j) `conv3*` (k) `conv4*` (l) `conv5*`

Figure 2: Hypercolumn features extracted from two sample images. Each map represents the mean activation map over a set of layers: (b) and (h) are built using layers `conv1_1` and `conv1_2`, (c) and (i) with layers `conv2_1` and `conv2_2`; (d) and (j) with `conv3_1`, `conv3_2` and `conv3_3`; (e) and (k) with `conv4_1`, `conv4_2`, and `conv4_3`. Finally, (f) and (i) are built using layers `conv5_1`, `conv5_2` and `conv5_3`. Best viewed in color.

of categories.

Each concept in WordNet is described by a set of words or word phrases (called *synonim set*, or *synset*). We match each unigram extracted from the text with the most semantic similar synset in a semantic space. In particular, we train a skip-gram model [16] on the dump of the English Wikipedia. The basic idea of skip-gram models is to fit the word embeddings such that the words in corpus can predict their context with high probability. Semantically similar words lie close to each other in the embedded space.

Word embedding algorithms assign each word to a vector in the semantic space, and the semantic similarity $S(u_1, u_2)$ of two concept terms u_1 and u_2 is defined as the cosine similarity between their word embeddings. For synsets, which do not have an explicit embedding, we take the average of the vectors from each word in the synset and L_2-normalize the average vector. The resulting similarity is used to match each concept with the nearest Imagenet category: given a unigram u found in text, the mapping function to the external corpus is as follows:

$$M(u) = \arg\max_{c \in \mathcal{C}} S(u, c) \tag{1}$$

where $\mathcal{C}$ is the set of all concepts in the corpus.

Having mapped each concept from the video collection to an external corpus, a classifier can be built to detect the presence of a visual concept in a shot. Since the number of terms mined from the text data is large, the classification step needs to be efficient, so instead of running the classifier on each frame of the video, we take the middle frame of each shot, using the shot detector in [1]. At the same time, given the temporal coherency of a video, it is unlikely for a visual concept to appear in a shot which is far from the point in which the concept found in the transcript. For this reason, we run a classifier only on shots which are temporally near to its corresponding term, and apply a Gaussian weight to each term based on the temporal distance.

Images from the external corpus are represented using feature activations from pre-trained CNNs. In particular, we employ the 16-layers model from VGG [20], pretrained on the ILSVRC-2012 [19] dataset, and use the activations from layer `fc6`. Then, a linear probabilistic SVM is trained for each concept, using randomly sampled negative images from other classes; the probability output of each classifier is then used as an indicator of the presence of a concept in a shot.

Formally, given a shot s which appears in the video at time t_s, and a unigram u found in transcript at time t_u, the probability that u is visually confirmed in s is computed as:

$$P(s, u) = f_{M(u)}(s) e^{-\frac{(t_u - t_s)^2}{2\sigma_a^2}} \tag{2}$$

where $f_{M(t)}(s)$ is the probability given by the SVM classifier trained on concept $M(t)$ and tested on shot s.

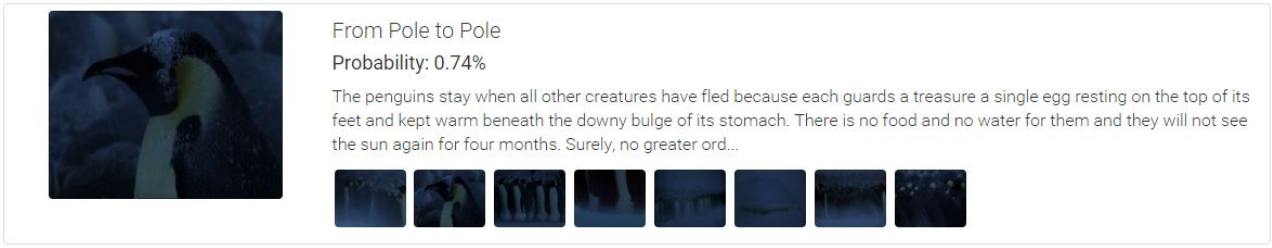
Aesthetic ranking The probability function defined above accounts for the presence of a particular visual concept in one shot, and is therefore useful to rank scenes given a user query. However, the thumbnail returned to the user should be visually representative as well. This requires a thumbnail selection step, which should account for low level characteristics, like color, edges and sharpness, as well as high level features, such as the presence of a clearly visible object in the center.

We claim that the need of low and high level features is an excellent match with the hierarchical nature of CNNs: convolutional filters, indeed, are known to capture low level as well as high level characteristics of the input image. This has been proved by visualization and inversion techniques, like [23] and [15], which can visualize the role of each filter.

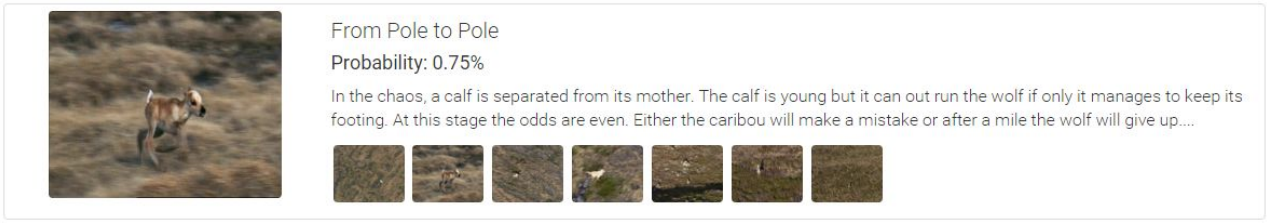
Being activations from convolutional filters discriminative for visual representativeness, a ranking strategy could be set up to learn their relative importance given a dataset of user preferences. However, medium sized CNNs, like the VGG-16 model [20], contain more than 4000 convolutional filters, which produce as much activation maps. This makes the use of raw activations infeasible with small datasets: moreover, maps from different layers have different sizes, due to the presence of pooling layers. To get around with this, we resize each activation map to fixed size with bilinear interpolation, and average feature maps coming from the different layers, inspired by the Hypercolumn approach presented in [8]. Moreover, since the the user usually focuses on the center of the thumbnail rather than its exterior, each maps is



Figure 3: Ranking of a sample scene. Thumbnails with a centered and clearly visible animal are preferred against blurred and low-quality frames (best viewed in color).



(a) Result for query *penguin*.



(b) Result for query *calf*.

Figure 4: Retrieval results. The same video is retrieved when searching for *penguin* and for *calf*, however, two different scenes are selected. Reported probability values correspond to $R_{scene}(q)$ in the paper.

multiplied by a normalized gaussian density map, centered on the center of the image and with standard deviation $\sigma_b \cdot l$, where $l \times l$ is the size of the CNN input.

Following the VGG-16 architecture, we build five hypercolumn maps, each one summarizing convolutional layers before each pooling layer: the first one is computed with activation maps from layers `conv1_1` and `conv1_2`; the second one with `conv2_1` and `conv2_2`; the third with `conv3_1`, `conv3_2` and `conv3_3`; the fourth with `conv4_1`, `conv4_2` and `conv4_3`; the last with `conv5_1`, `conv5_2` and `conv5_3`. An example of the resulting activation maps is presented in Fig. 2: as it can be seen, both low level and high level layers are useful to distinguish between a significant and non significant thumbnail.

To learn the relative contribution of each hypercolumn map, we rank thumbnails from each scene according to their visual representativeness, and learn a linear ranking model. Given a dataset of scenes $\{s_i\}_{k=0}^{n}$, each with a ranking r_k^*, expressed as a set of pairs (d_i, d_j), where thumbnail d_i is annotated as more relevant than thumbnail d_j, we minimize the following function:

$$\begin{aligned} \underset{\mathbf{w},\epsilon}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i,j,k}\epsilon_{i,j,k} \\ \text{subject to} \quad & \forall (d_i,d_j) \in r_1^* : \mathbf{w}\phi(d_i) \geq \mathbf{w}\phi(d_j) + 1 - \epsilon_{i,j,1} \\ & \dots \\ & \forall (d_i,d_j) \in r_n^* : \quad \mathbf{w}\phi(d_i) \geq \mathbf{w}\phi(d_j) + 1 - \epsilon_{i,j,n} \\ & \forall i,j,k : \epsilon_{i,j,k} \geq 0 \end{aligned} \tag{3}$$

where $\phi(d_i)$ is the feature vector of thumbnail d_i, which is composed by the mean and standard deviation of each hypercolumn map extracted from the thumbnail itself. C allows trading-off the margin size with respect to the training error. The objective stated in Eq. 3 is convex and equivalent to that of a linear SVM on pairwise difference vectors $\phi(d_i) - \phi(d_j)$ [9].

Retrieval Given a query q, we first match q with the most similar detected concept u, using the Word2Vec embedding. If the query q is composed by more than one words, the mean of the embedded vectors is used. Each scene inside the video collection is then assigned a score according to the following function:

$$R_{scene}(q) = \max_{s \in scene} \left(\alpha P(s,u) + (1-\alpha) \max_{d \in s} \mathbf{w}\phi(d) \right) \tag{4}$$

where s is a shot inside the given scene, and d represent all keyframes extracted from a given shot. Parameter α tunes the relative importance of semantic representativeness and aesthetic beauty. The final retrieval results is a collection of scenes, ranked according to $R_{scene}(q)$, each one represented with the keyframe that maximizes the second term of the score.

From an implementation point of view, $P(s,u)$ can be computed offline for each unigram u found in the transcript, for example with an inverted index. $\mathbf{w}\phi(d)$, as well, can be computed in advance for each key-frame, thus greatly reducing the computational needs in the on-line stage.

Episode	Shots	Scenes	Unigrams
From Pole to Pole	450	66	337
Mountains	395	53	339
Fresh Water	425	62	342
Caves	473	71	308
Deserts	461	65	392
Ice Worlds	529	65	343
Great Plains	534	63	336
Jungles	418	53	346
Shallow Seas	368	62	370
Seasonal Forests	393	57	356
Ocean Deep	470	55	333

Table 1: Statistics on the *BBC Planet Earth* dataset.

4. EXPERIMENTAL RESULTS

We evaluate the proposed method on a collection of 11 episodes from the *BBC Planet Earth*[2] series. Each video is approximately 50 minutes long, and the whole dataset contains around 4900 shots and 670 scenes. Each video is also provided with the transcript, and on the whole dataset a total of 3802 terms was extracted using the CoreNLP parser. Table 1 reports some statistics on the dataset. Parameters σ_a and σ_b were set to 5 and 4.5 respectively, while C was set to 3.

4.1 Thumbnail selection evaluation

Since aesthetic quality is subjective, three different users were asked to mark all keyframes either as aesthetically relevant or non relevant for the scene they belong to. For each shot, the middle frame was selected as keyframe. Annotators were instructed to consider the relevance of the visual content as well as the quality of the keyframe in terms of color, sharpness and blurriness. Each keyframe was then labeled with the number of times it was selected, and a set of (d_i, d_j) training pairs was built according to the given ranking, to train our aesthetic ranking model.

For comparison, an end-to-end deep learning approach (*Ranking CNN*) was also tested. In this case the last layer of a pre-trained VGG-16 network was replaced with just one neuron, and the network was trained to predict the score of each shot, with a Mean Square Error loss. Both the Ranking CNN model and the proposed Hypercolumn-based ranking were trained in a leave-one-out setup, using ten videos for training and one for test.

Table 2 reports the average percent of swapped pairs: as it can be seen, our ranking strategy is able to overcome the Ranking CNN baseline and features a considerably reduced error percentage. This confirms that low and high level features can be successfully combined together, and that high features alone, such as the ones the Ranking CNN is able to extract from its final layers, are not sufficient. Figure 3 shows the ranking results of a sample scene: as requested in the annotation, the SVM model preferred thumbnails with good quality and a clearly visible object in the middle.

4.2 Retrieval results evaluation

On a different note, we present some qualitative results of the retrieval pipeline. Figure 4 shows the first retrieved result when searching for *penguin* and *calf*, using $\alpha = 0.5$. As it can be seen, our method retrieves two different scenes for the same video, based on the visually confirmed concepts extracted from the transcript. The same video, therefore, is presented with different scenes depending on the query. Moreover, selected thumbnails are actually representative of the query and aesthetically pleasant, when compared to the others available keyframes for those scenes. Depending on the query, it may also happen that the same scene is presented with two different thumbnails, as depicted in Fig. 5: in this case the same scene was retrieved with query *ant* and *spider*; however, in the first case the selected thumbnail actually represents an ant, while in the second case a spider is selected, thus enhancing the user experience.

[2] `http://www.bbc.co.uk/programmes/b006mywy`

Episode	Ranking CNN	Hypercolumns Ranking
From Pole to Pole	8.23	4.10
Mountains	12.08	7.94
Fresh Water	12.36	8.11
Caves	9.98	8.76
Deserts	13.90	9.35
Ice Worlds	6.62	4.33
Great Plains	10.92	9.63
Jungles	12.28	7.43
Shallow Seas	10.91	6.22
Seasonal Forests	9.47	4.82
Ocean Deep	10.73	5.75
Average	**10.68**	**6.95**

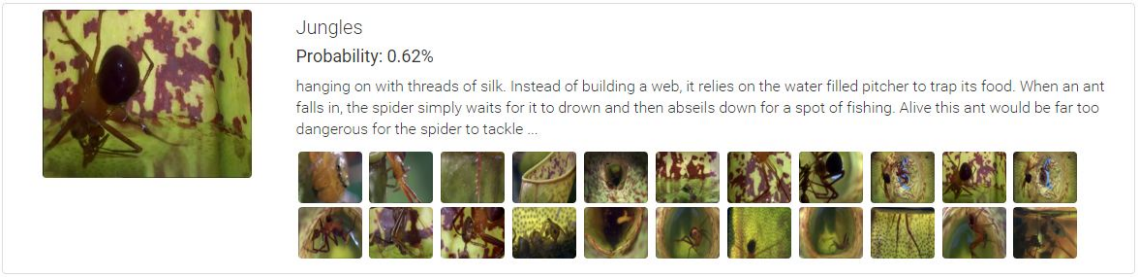
Table 2: Aesthetic ranking: average percent of swapped pairs on the *BBC Planet Earth* dataset (lower is better).

4.3 User evaluation

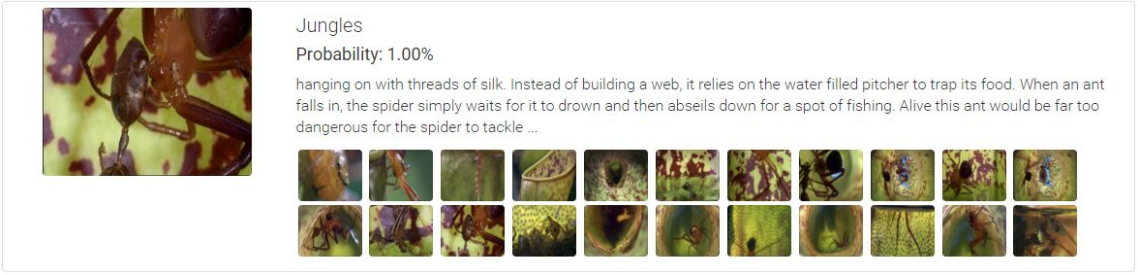
To quantitatively evaluate the ranking results and their effect on user experience, we conducted a user study with 12 undergraduate students. A demonstration and evaluation interface was built, in which the first three results returned by our method could be directly compared with three scenes retrieved with a full-text search inside the transcript, and presented with a random thumbnail different from the one selected by our system. As in Fig. 4 and 5, each retrieved scene was presented with the selected thumbnail, the corresponding transcription and with all the key-frames extracted from the scene. Users could also click on the thumbnail to watch the corresponding scene.

Evaluators were asked to compare the provided result sets and vote the one they liked most, for a set of 20 queries. Results from our method were preferred to those provided by the baseline in the 82% of cases, in the 15% of evaluations they were said to be equivalent, while in the remaining 3% of cases the baseline was preferred. The same queries were presented again replacing the thumbnails selected by our method with random ones. In this case the preferences were 12% for the baseline and 57% for our proposal, while in the 31% of cases results were evaluated as equivalent.

This confirms the role of selecting appropriate thumbnails when dealing with casual users (the students didn't have any real goal, nor were particularly interested in the queries we provided). One of the conclusions we can draw from this tests is that the presentation of the results may strongly influence the feeling of "correctness" of the retrieved results.



(a) Result for query *ant*.



(b) Result for query *spider*.

Figure 5: Retrieval results. In this case the same scene from the same video is retrieved with two different queries (*ant* and *spider*), however, two different (and significant) thumbnails are selected. Reported probability values correspond to $R_{scene}(q)$ in the paper.

5. CONCLUSIONS

We presented a novel video retrieval pipeline, in which videos are decomposed into short parts (namely scenes), that are used as the basic unit for retrieval. A score function was proposed to rank scenes according to a given textual query, taking into account the visual content of a thumbnail as well as its aesthetic quality, so that each result is presented with an appropriate keyframe. Both the semantics and the aesthetics were assessed using features extracted from Convolutional Neural Networks, and by building on-the-fly classifiers for unseen concepts. Our work has been evaluated both in qualitative and quantitative terms, and results in enhanced retrieval results and user experience.

Acknowledgments We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research. This work was carried out within the project "Città educante" (CTN01_00034_393801) of the National Technological Cluster on Smart Communities cofunded by the Italian Ministry of Education, University and Research - MIUR.

6. REFERENCES

[1] E. Apostolidis and V. Mezaris. Fast Shot Segmentation Combining Global and Local Visual Descriptors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6583–6587, 2014.

[2] L. Ballan, M. Bertini, G. Serra, and A. Del Bimbo. A data-driven approach for tag refinement and localization in web videos. *Computer Vision and Image Understanding*, 140:58–67, 2015.

[3] L. Baraldi, C. Grana, and R. Cucchiara. A deep siamese network for scene detection in broadcast videos. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, MM '15, pages 1199–1202, New York, NY, USA, 2015. ACM.

[4] B. Craggs, M. Kilgallon Scott, and J. Alexander. Thumbreels: query sensitive web video previews based on temporal, crowdsourced, semantic tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1217–1220. ACM, 2014.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[6] M.-C. De Marneffe, B. MacCartney, C. D. Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[9] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[10] H.-W. Kang and X.-S. Hua. To learn representativeness of video frames. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 423–426. ACM, 2005.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[12] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2(1):1–19, 2006.

[13] C. Liu, Q. Huang, and S. Jiang. Query sensitive dynamic web video thumbnail generation. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2449–2452. IEEE, 2011.

[14] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3715, 2015.

[15] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 5188–5196. IEEE, 2015.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[17] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[18] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 435–441. IEEE, 2006.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, pages 1–42, April 2015.

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

[21] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[22] C. G. Snoek, B. Huurnink, L. Hollink, M. De Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *Multimedia, IEEE Transactions on*, 9(5):975–986, 2007.

[23] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer, 2014.