NeuralStory: an Interactive Multimedia System for Video Indexing and Re-use

NeuralStory: an Interactive Multimedia System for Video Indexing and Re-use

http://neuralstory.ing.unimore.it

Lorenzo Baraldi, Costantino Grana and Rita Cucchiara
Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Email: {name.surname}@unimore.it

***Abstract*—In the last years video has been swamping the Internet: websites, social networks, and business multimedia systems are adopting video as the most important form of communication and information. Video are normally accessed as a whole and are not indexed in the visual content. Thus, they are often uploaded as short, manually cut clips with user-provided annotations, keywords and tags for retrieval. In this paper, we propose a prototype multimedia system which addresses these two limitations: it overcomes the need of human intervention in the video setting, thanks to fully deep learning-based solutions, and decomposes the storytelling structure of the video into coherent parts. These parts can be shots, key-frames, scenes and semantically related stories, and are exploited to provide an automatic annotation of the visual content, so that parts of video can be easily retrieved. This also allows a principled re-use of the video itself: users of the platform can indeed produce new storytelling by means of multi-modal presentations, add text and other media, and propose a different visual organization of the content. We present the overall solution, and some experiments on the re-use capability of our platform in edutainment by conducting an extensive user evaluation with students from primary schools.**

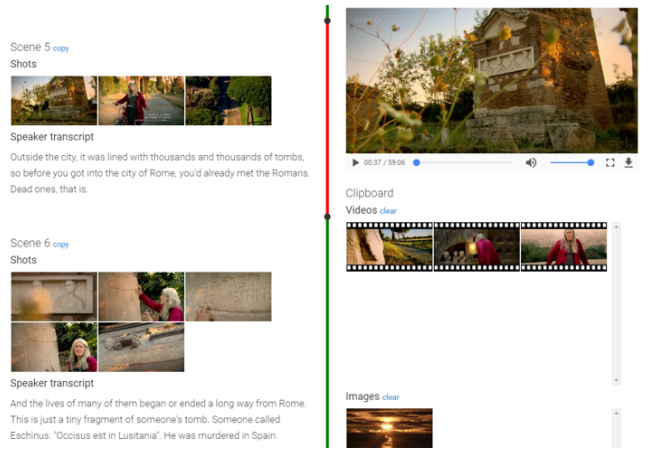


Fig. 1. Screenshot of the interface. In *NeuralStory*, video clips are automatically segmented in coherent parts to ease the browsing of long video clips. The interface also allows the user to pick selected and segments and re-use them to build multi-modal presentations.

I. Introduction

Video has become the largest source of traffic on the Internet: initially used for fun and entertainment, it is nowadays employed for social, commercial and business motivations. Video is expected by Cisco to account for 80% of all Internet traffic by 2019 [1], and Facebook stated that 90% of the social network content will be video-based by 2018. Accordingly, the Web Video Marketing Council reported that online video has become a crucial part in the 96% of the sales and marketing programmes for most business-to-business organisations [2]. In education, children and students make use of video as an enjoyable source of knowledge.

Unfortunately, re-using existing video footage is cumbersome. Commercial video browsing platforms like Youtube, Vimeo and Dailymotion treat the video as a indivisible entity, so that the user, after having found a video, receives no help in finding the portion of the video that really interests him. The user must either watch the entire video or move from one portion of the video to another through seek operations. This paradigm works well for short user generated video with a mostly uniform content, but not for longer edited video, which usually contain complex story lines with different topics. Consider for instance long documentary video clips, where only short segments are focused on a specific topic of a school lecture, or a long movie, where only a given scene may be of interest to create an advertisement, or to enforce an emotional concept in professional video production.

Retrieval is normally carried out by exploiting annotations provided by the uploader (such as the title, a description, and tags), and therefore relies on the hypothesis that user-generated annotations are complete and consistent. This is very often not true: the visual content is very rarely annotated with tags or natural language sentences, especially for long and edited video clips, since annotating every possibly relevant entity would be frustrating and time consuming. Finally, there is no direct link between the available annotation and the precise part of the video where an entity appears in the video, and there is no support to provide users with the portion of the video matches a query.

In this paper, we present a video browsing and retrieval platform which tries to move a step forward by addressing these limitations, and reducing the need of user-provided annotations. In our system, the storytelling structure of the video is automatically decomposed into shorter coherent segments by means of machine learning, so that users can easily get a glance of the content and retrieve specific parts of video.

A deep learning architecture has been trained using images, audio, speech-to-text, and textual content to segment the video into stories and to annotate connected key-frames with tags extracted by the visual content and by the speech dialogues.

The decomposition of a video also enables the re-use of the video itself: users, once provided with suitable fine-grained retrieval tools, can indeed use parts of different video clips to create a new multi-modal product. To showcase this advantage, our system lets the user pick video segments or frames to build presentations with multimedia slides, which we name *MeSlides*. This can became a valuable tool for any type of re-use: commercial marketing for business employers and publishers, digital humanity and cultural service for scholars, edutainment services for teaching assistance and schoolworks or homeworks. We evaluated the features of the system in this last use case scenario.

As well, the platform provides a summarization service, since it allows to show the storytelling of a video with a limited sequence of scene keyframes, by summarizing a hour-long video in a few seconds.

Since the underlying intuition of the system is that of decomposing the storytelling structure of the video to eventually create new storytelling, and most of the algorithms are based on Deep Learning, we name our platform *NeuralStory*. An interactive demo of the system is publicly available at the address provided in the heading of the page.

II. Related works

One of the main features of the system is temporal video decomposition, which has been known in literature as *scene detection*. This topic has been extensively addressed with method based on empirical rules [3], [4], on graph analysis [5] and on clustering [6]. The focus of previous works has been that of detecting location changes in movies: in our work, we go a step beyond, and employ a temporal video segmentation approach we developed [7] to identify topic changes in video clips rather than simple location changes. This is particularly beneficial in the case of documentaries and educational video, where the expected decomposition must necessarily rely on higher level semantic features.

On a different note, our system is also endowed with a retrieval component. This is, indeed, a topic in which lot of work has been proposed. Since the seminal *Video Google* [8], which was based on bag-of-words matching, concept-based methods have emerged as a popular approach to video retrieval. Snoek *et al.* [9] 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 [10], 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.

Browsing and visualization interfaces for video have also been widely studied. Many researchers investigated the benefit of replacing still images with moving thumbnails, i.e., small clips extracted from the video. [11] proposed an aligned thumbnails-based video browsing system using content-based video browsing approach by automatically detecting scenes. Their work exploits a hierarchical sliding interface to browse the playing video. Jiang *et al.* [12] addressed the problem of browsing the results of a complex query, by building a hierarchical visualization which can help users quickly understand multiple facets of a query topic in an organized manner, by using the hierarchy of textual descriptions available on Wikipedia.

III. Temporal Video Segmentation

Decomposing a video into coherent parts for video re-use is an intrinsic multi-modal task; it cannot be solved by applying heuristic rules, or a-priori defined models due to the variety of boundaries. They can be due either to a discontinuity of visual data either to a stop in music and audio, either in a change of speech description. For this reason, we employ a combination of multi-modal features and a supervised clustering algorithm in a Triplet Deep Network, thus learning from examples annotated by different users the optimal way of segmenting the video [7]. In the following, we will employ the term *scene* when referring to such coherent video parts, as this is the term traditionally used in literature.

The video is firstly decomposed into a set of chunks taken by the same camera (*i.e.* shots), using an open source shot detector [13]. The level of granularity given by shots is far more subtle than the one we want to obtain, nevertheless, since the content of a shot is usually uniform, we can constrain scene boundaries to be a subset of shot boundaries, therefore reducing the problem of scene detection to that of clustering adjacent shots. This preliminary decomposition also reduces the computational efforts needed to process the entire video, given that few key-frames can be used as the representative of the whole shot. Similarly, features coming from other modalities (like text transcript, audio and semantics) can be encoded at the shot level by following the same homogeneity assumption.

For each shot of the video, we extract the following four different features, in order to take into account all the modalities present in the video.

1) Visual appearance: We encode the visual appearance of a shot by means of a pre-trained Convolutional Neural Network (namely, the VGG-16 model [14]) which is shortened by cutting out the last fully connected layers. This extracts high level features from the input image, which can be a rich source of information to identify changes in visual content between one portion of the video and another. Given that a single key-frame might be too poor to describe a shot, we uniformly sample three key-frame from the input shot, and take the pixel-wise maximum of the network responses.

2) Audio features: Audio is another meaningful cue for detecting scene boundaries, since audio effects and soundtracks are often used to underline the development of a scene or a change in content. We extract MFCCs descriptors [15] over a

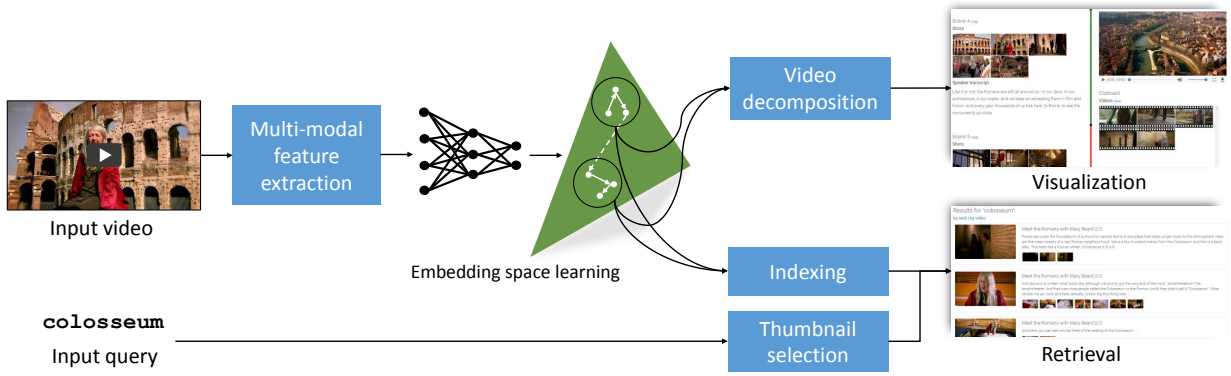


Fig. 2. Overview of the video annotation and retrieval pipeline. Our system decomposes the input video into coherent parts by means of a Triplet Deep Network trained on multi-modal features: this decomposition is the basis of the visualization interface, and also allows a fine-grained search inside video clips. Retrieval is carried out by leveraging automatic annotation and a thumbnail selection strategy, so that search results are both semantically and esthetically relevant.

10ms window. The MFCC descriptors are aggregated by Fisher vectors using a Gaussian Mixture Model with 256 components.

3) Quantity of Speech: A pause in the speaker discourse can be relevant to identify a change of scene: for this reason, we turn to the video transcript and build a feature which computes the amount of words being said inside a shot. The quantity of speech is computed as the number of words which appear in that shot, normalized with respect to the maximum number of words found in a shot for the full video.

4) Textual Concept feature: Beside visual appearance and audio, we also want to encode the semantics of the transcript of the video. Using a part-of-speech tagger [16], we select unigrams which are annotated as noun, proper noun and foreign word, which are more likely to be representative of the underlying topic. Unigrams from the whole video collection are then clustered according to their semantics, by projecting them to a semantic space [17] trained on the dump of the English Wikipedia, and using Spectral Clustering to detect K different concept groups.

Concept groups provide an ideal mean to describe topic changes in text. A textual feature vector, $\mathbf{t}(s)$, is built to get a representation of how much each concept group is present in a shot s and in its neighborhood

$$\mathbf{t}(s) = \left[\sum_{t \in \mathcal{T}_1} e^{-\frac{(u_t - u_s)^2}{2\sigma^2}}, \sum_{t \in \mathcal{T}_2} e^{-\frac{(u_t - u_s)^2}{2\sigma^2}}, ..., \sum_{t \in \mathcal{T}_K} e^{-\frac{(u_t - u_s)^2}{2\sigma^2}} \right] \tag{1}$$

where $\mathcal{T}_k$ is the set of all appearances in the video of terms belonging to a cluster (in which, therefore, some terms may appear multiple times), u_t is the time in which term t appears, and u_s is the time of the central position of shot s. Practically, the summations in Eq. 1 are approximated by considering a sufficiently large neighborhood of the shot, instead of using the entire video.

5) Visual Concept feature: Many of the terms in the transcript may refer to abstract concepts, which do not appear visually in the video. We are therefore interested into defining a second feature vector to account for this scenario, by visually confirming concepts which are found in the transcript.

Firstly, we build a connection between the textual and the visual domain by matching each unigram with the most similar Imagenet [18] class, by means of the same semantic embedding space [17]. Using this mapping we then build a classifier to detect the presence of a textual concept in a shot. Images from the external corpus are represented using feature activations from the pre-trained VGG-16 model. Then, a linear SVM is trained for each concept, using randomly sampled negative training data; the probability output of each classifier is then used as an indicator of the presence of a concept in a shot.

Formally, the visual concept feature of shot s, $\mathbf{v}(s)$, is a K-dimensional vector, defined as

$$\mathbf{v}(s) = \left[\sum_{t \in \mathcal{T}_1} f_t(s) \cdot e^{-\frac{(u_t - u_s)^2}{2\sigma^2}}, ..., \sum_{t \in \mathcal{T}_K} f_t(s) \cdot e^{-\frac{(u_t - u_s)^2}{2\sigma^2}} \right] \tag{2}$$

where $f_t(s)$ is the probability given by the SVM classifier trained on term t and tested on shot s.

The overall feature vector for a shot is the concatenation of visual and textual concept features vector and of visual appearance and quantity of speech feature vectors. To this, we also add information about the timestamp and the length of a given shot. A Triplet Deep Network is then trained on ground-truth decompositions by minimizing a contrastive loss function. In particular, the multi-modal feature vector of a shot, $\mathbf{x}_i$, is fed to a series of three fully connected trainable layers (with, respectively, 500, 125 and 30 neurons), to compute an embedding function $\phi(\mathbf{x}_i)$. We then sample a triplet of shots $(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-)$, where $\mathbf{x}_i^+$ is constrained to belong to the same scene of $\mathbf{x}_i$, and $\mathbf{x}_i^-$ is constrained to belong to a different scene. The following cost function is then computed for each training triplet:

$$\max\left(0, \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_i^+)\|^2 + \left(1 - \| \phi(\mathbf{x}_i) - \phi(\mathbf{x}_i^-) \|^2\right)\right). \tag{3}$$

Optimization is carried out trough mini-batch Stochastic Gradient Descent.

At test time, the network has learned to distinguish similar and dissimilar shots, and can be therefore employed to perform temporal video segmentation. In particular, our clustering algorithm relies on the minimization of the total within-group sum of squares (TWSS) inside each scene [7]. Differently from k-means, we would also like to find the number of clusters, with the additional constraint of them being temporally continuous intervals. Minimizing the TWSS alone would lead to the trivial solution of having a single shot in each sequence, so a penalty term needs to be added to avoid over-segmentation. The objective we solve is thus

$$\min_{m,t_1,\ldots,t_m} \sum_{i=0}^{m} WSS_{t_i,t_{i+1}} + Cg(m,n) \tag{4}$$

where $g(m,n) = m(\log(n/m)+1)$ is a Bayesian information criterion penalty.

A. Visualization interface

The decomposition of a video into coherent parts allows a better visualization of the video, in which the user can get a glance of the content without having to watch the entire video, and also jump from one point to another, given that the decomposition intrinsically allows a summarization of the video.

In the interface (see Figure 1 for an example), we place the classical video player on one side of the page, and keep the left-most side to show the decomposition of the video in a timeline fashion. For each scene, we show a summary of the key-frames, the transcript, and detected visual concepts. Users can scroll the timeline to get a detailed view of the content a video in seconds: key-frames provide a visual preview, while transcript and visual concepts give a quick insight of what the speaker is talking about.

IV. Indexing and Search

The ability to index parts of a video is an essential feature of the platform, as it enables a fine-grained search which is also important for video re-use. In developing this feature, we considered three main objectives. First of all, video clips should be indexed at the scene level, so that users can search inside video clips and not only among video clips. Secondly, we wanted the user to be very confident on the relevance of a result just by looking at the provided thumbnail: this led us to rethink the way thumbnails are chosen, and to identify a strategy to select query-depend and aesthetically valuable thumbnails. Lastly, we constrained the indexing system to be fully automatic, and therefore chose to rely on concept detection algorithms solely, rather than exploiting user-generated annotations.

In *NeuralStory* indexing is performed by re-using the classification step used to compute the visual concept feature (see Section III-5). For each detected scene, indeed, we store the probability that each of the terms present in the transcript is actually present in the scene, relying on visual classifiers built on-the-fly for each unigram found in the transcript. From an implementation point of view, we use an hash table indexed with unigrams, so that performing a query for a word which appears in the transcript requires, on average, $O(1)$. For each key, we store an array of pointers to scenes, with all the scenes in which the word appears, along with the probability of each key-frame of the scene. To limit the size of the index, we only store scenes with at least one key-frame with probability greater than a threshold. This is not critical and for our experiments it has been fixed to 0.3.

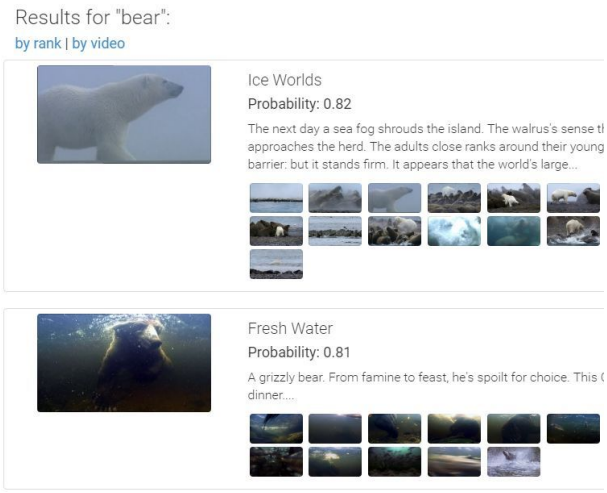


Fig. 3. Retrieval interface for query *bear*. Differently from commercial video browsing platforms, each result represents a portion of a video (i.e. a *scene*), and thumbnails are selected according to the input query and aesthetic criteria.

To tackle the case in which the query performed by the user does not correspond to a unigram, we also maintain an array with all the keys of the hash table (i.e. all the unigrams). We then perform a nearest neighbour matching using the semantic space of [17] between the query and all unigrams available. In case the query contains multiple words, we take the average L2-normalized vector as its representation, and perform the nearest neighbour matching as usual.

On a different note, we also wanted to include a thumbnail selection strategy in our system, which could provide the user with query-dependent and aesthetically valuable key-frames. For this reason, we follow the approach we presented in [19]: a linear SVM ranking model is built using features extracted from multiple levels of a CNN, and trained using aesthetically relevant and non relevant pairs. The score given by the aesthetic ranker is then merged with that of the visual classifier to select the best key-frame for the given query.

Figure 3 shows the retrieval interface for query "bear". As it can be seen, the interface provides a list of results, each representing a single scene, along with the title of the video and the transcript of the scene. Thumbnails are selected, among the set of available key-frames, according to the semantics of the specific query and their aesthetic relevance.

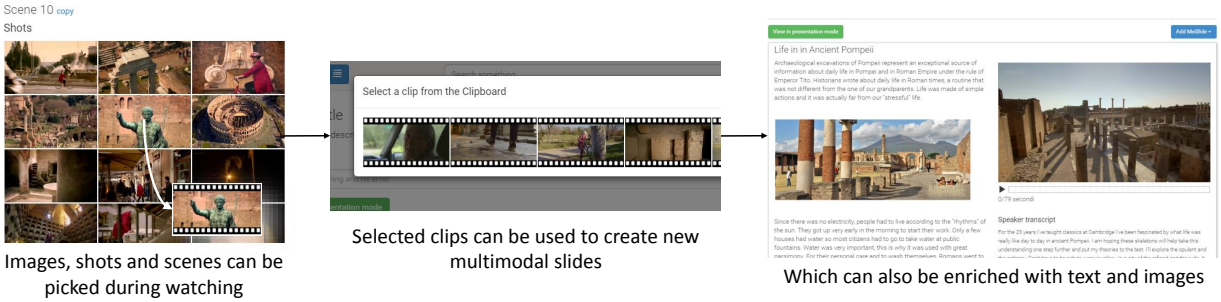


Fig. 4. Video re-use pipeline in *NeuralStory*. Users can pick images, shots and video clips while surfing the video collections. Selected images and fragments can then re-used to build new multi-modal presentations.

V. Enabling re-use through MeSlides

The decomposition of video clips into scenes and the availability of a scene-based retrieval system enable the re-use of parts of video inside the platform. To showcase this re-use capability, we endowed the system with a strategy to select relevant video parts, and an application capable of building multi-modal presentations using the selected material.

When watching or searching a video the user can select scenes, shots or even key-frames and drop them into a clipboard for further re-use (see Figure 4 for an example). In a second section of the platform, instead, it is possible to build presentations with multimedia slides (or *MeSlides*, as we call them), in which users can insert text, images, and parts of video dragged in the clipboard. All the content of a slide is dynamically editable by simply clicking on a specific region of the slide itself, and modifications are automatically stored server-side with AJAX calls.

Each presentation has a title, a short description, and can be annotated with a set of tags. Users can create four types of MeSlides, depending on the content they would like to add: slides with only text on one column, with text on two columns, with only a video and with text and video on two columns. Frames and video clips picked in the clipboard, as well as images coming from Google Images search, can be added without leaving the page.

From an implementation point of view, the playing of a specific part of a video is realized with a Javascript modification of the default HTML5 video player, which makes it possible to play a time interval of a video without having to cut the video file on the server side. This is particularly beneficial, as we do not need the permission to modify the original video file from the content producer. In other words, the decomposition carried out by the system, in all perspectives, does not require to actually modify the original video, and thus does not require any special right on the video, except from that of reproducing it publicly.

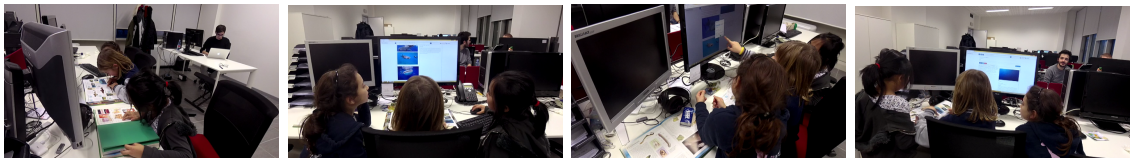
VI. User evaluation

To evaluate the features of the system, we defined a user study in a setting in which the re-use of existing footage could be particularly beneficial. Given the progressive modernization of schools and teaching paradigms, primary and junior high school are a suitable environment for the test. The retrieval and re-use capabilities of our system are indeed a perfect match with the need of students and teachers to find relevant video content for building researches and presentations.

We uploaded on the system a broadcast video collection, coming from the *BBC Planet Earth* and *BBC Romans with Mery Beard* series. This prototype database consists, in total, of 19 video clips, with average length of 50 minutes, which are particularly suitable for our use case, given their length and variety of contents. Table I reports the number of shots, scenes and searchable unigrams for each video of the database. As it can be seen, the variety of unigrams enables a principled user evaluation of the concepts behind our framework.

Two kind of user studies have been conducted. Firstly, we tested the system and the prototype interface with small groups of students from primary schools in a controlled environment (Figure 5a). In this study, children were introduced to the interface by an adult, and then asked to compose a short schoolwork with MeSlides, on a topic of their choice. Children were given school books, and were instructed to familiarize with the topic by searching on the books and by using the retrieval interface. During the process, an adult was present to assist them and take note of their difficulties in using the system. This, later, gave valuable information on how to enhance the usability of the interface and the process of building slides.

A more extensive evaluation was then performed by involving four primary classes from two different schools. Some pictures taken during the study are reported in Figure 5b. In this case, tests were performed with the whole class, and in a real environment. Teachers and children were instructed on how to use the interface during a 30 minutes introduction, in which a sample presentation was created by the instructor taking into account ideas and suggestions from the students. Later, students were organized in pairs, and an Android tablet was given to each pair, asking them to create a presentation on a topic of their choice. At the end of the user study, students were able to produce a total of 65 presentations, with one or more slides, and without encountering any substantial difficulty during the process. This confirms the usability of the interface, and that the interface is actually suitable for assisting students and teachers in producing schoolworks.



(a) Prelimiary user tests in the lab, with three children

(b) Extensive user evaluation in four primary classes

Fig. 5. User study with primary school students.

TABLE I
STATISTICS ON THE *BBC Planet Earth* AND *BBC Romans with Mery Beard* DATASETS.

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
Meet the Romans 1/3	856	114	746
Meet the Romans 2/3	684	105	800
Meet the Romans 3/3	673	68	732
Empire Without Limit 1/4	773	112	730
Empire Without Limit 2/4	809	111	779
Empire Without Limit 3/4	825	139	706
Empire Without Limit 4/4	696	134	707
Pompeii	675	93	638

VII. CONCLUSION

We presented a novel system for video browsing and retrieval, which also allows a principled re-use of existing footage. The proposal connects temporal video segmentation and automatic annotation algorithms, together with a visualization and re-use interface. By means of NeuralStory, the storytelling structure of edited video clips is decomposed, to eventually build new multi-modal artifacts in the form of presentations. An extensive user evaluation has been performed, by involving children of four primary classes, to assess the usefulness and usability of the system.

ACKNOWLEDGMENT

Our work is partially founded by 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). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

REFERENCES

[1] I. Cisco Systems, "Cisco VNI Forecast and Methodology, 2015-2020," Tech. Rep., 2015.
[2] Web Video Marketing Council, "Video Statistics: The Marketer's Summary 2016," Tech. Rep., 2016.
[3] C. Liu, D. Wang, J. Zhu, and B. Zhang, "Learning a Contextual Multi-Thread Model for Movie/TV Scene Segmentation," *IEEE TMM*, vol. 15, no. 4, pp. 884–897, 2013.
[4] V. T. Chasanis, C. Likas, and N. P. Galatsanos, "Scene detection in videos using shot clustering and sequence alignment," *IEEE TMM*, vol. 11, no. 1, pp. 89–100, 2009.
[5] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE TMM*, vol. 7, no. 6, pp. 1097–1105, 2005.
[6] A. Hanjalic, R. L. Lagendijk, and J. Biemond, "Automated high-level movie segmentation for advanced video-retrieval systems," *IEEE TCSVT*, vol. 9, no. 4, pp. 580–588, 1999.
[7] L. Baraldi, C. Grana, and R. Cucchiara, "Recognizing and presenting the storytelling video structure with deep multimodal networks," *IEEE TMM*, 2017.
[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. IEEE, 2003, pp. 1470–1477.
[9] C. G. Snoek, B. Huurnink, L. Hollink, M. De Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE TMM*, vol. 9, no. 5, pp. 975–986, 2007.
[10] L. Ballan, M. Bertini, G. Serra, and A. Del Bimbo, "A data-driven approach for tag refinement and localization in web videos," *Comput. Vis. Image Und.*, vol. 140, pp. 58–67, 2015.
[11] U.-N. Yoon, M.-H. Ga, and G.-S. Jo, "Aligned thumbnails-based video browsing system with chromecast," in *Consumer Electronics (ICCE), 2015 IEEE International Conference on*. IEEE, 2015, pp. 86–87.
[12] Y.-G. Jiang, J. Wang, Q. Wang, W. Liu, and C.-W. Ngo, "Hierarchical visualization of video search results for topic-based browsing," *IEEE TMM*, vol. 18, no. 11, pp. 2161–2170, 2016.
[13] L. Baraldi, C. Grana, and R. Cucchiara, "Shot and scene detection via hierarchical clustering for re-using broadcast video," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2015, pp. 801–811.
[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[15] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, 2000.
[16] M.-C. De Marneffe, B. MacCartney, C. D. Manning *et al.*, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, 2006, pp. 449–454.
[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *ANIPS*, 2013, pp. 3111–3119.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.
[19] L. Baraldi, C. Grana, and R. Cucchiara, "Scene-driven retrieval in edited videos using aesthetic and semantic deep features," *ACM International Conference on Multimedia Retrieval*, 2016.