A Deep Look Into Educational Videos Indexing

Junjie Cai, Michele Merler, *Member, IEEE*, Sharath Pankanti, *Fellow, IEEE*, and Qi Tian, *Member, IEEE*

***Abstract*—In the last decade, the explosive growth of online learning videos has gathered a wealth of attention from both industry and academia communities. With such a large scale of educational content on the Web, how to automatically index and organize these videos has become an interesting and challenging task. Moreover, there exist a large quantity of educational videos containing a variety of visual content ranging from natural landscapes to classrooms, from animations to kids, from animals to plants, *etc*. This constitutes of a special class of educational videos which thus far has not been investigated under the perspective of visual indexing, since all efforts on educational content indexing have been focused on more traditional lecture style videos.**

Inspired by the recent progress of semantic-based approaches on indexing unconstrained consumer videos, we investigate the performance of semantic visual classifiers applied to the educational video data domain. In this paper, we propose to build a set of visual classifiers targeting at a real-world dataset of educational videos. which contains 370 clips of videos with more than 30 thousands key frames provided from non-lecture education collection. Moreover, to enhance the discriminative power of the proposed classifiers by complementing traditional low level descriptors, we leverage state-of-the-art deep learning features and test them in the educational video indexing domain, as opposed to the traditional one of natural images classification. Extensive retrieval experiments on two real-world datasets of educational videos using our visual content classifiers demonstrate the effectiveness of the proposed approach assisted by deep features.

***Index Terms*—Deep Learning, Educational Video Indexing.**

I. Introduction

In Web 2.0 era, online education is a popular way for students to acquire novel knowledge in a self-paced and asynchronous way. In the last decade, the explosive growth of online learning videos has enabled people to connect, express their ideas, and share interests [4]. Popular online learning websites including Coursera[1], Udacity[2], MIT Open CourseWare[3]and Udemy[4] regularly post instructors' lectures as video clips for the fruition of a wider audience. Sites like Videolectures[5] work as public venues where recorded author's academic presentation videos are shared. TED[6] and YouTube EDU[7] are popular for sharing and expressing novel technology ideas. Youtube also features channels with education specific content such as SciShow, STEMBite and CrashCourse. Many companies run corporate training via instructional videos, and traditional publishers of educational content such as Pearson[8] are producing increasingly more video content to complement and/or substitute traditional media. Some exemplar educational videos and key frames extracted from them are illustrated in Figure 1.

Fig. 1. Example of educational online learning videos and extracted frames.

With such a great number of educational videos gathered from the Web, automatically classifying and indexing those videos is an interesting and challenging task. To date, many state-of-the-art online learning platforms leverage text retrieval techniques to search for relevant educational videos. Specifically, text-based

J. Cai and Q. Tian are with the University of Texas at San Antonio, 78256, USA (email: caijunjieustc@gmail.com).

M. Merler and S. Pankanti is with IBM Thomas J Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY, 10598

[1]http://www.coursera.org
[2]http://www.udacity.com/
[3]http://ocw.mit.edu
[4]http://www.udemy.com/
[5]http://videolectures.net/
[6]http://www.ted.org
[7]http://www.youtube.com/edu
[8]http://www.pearson.com/

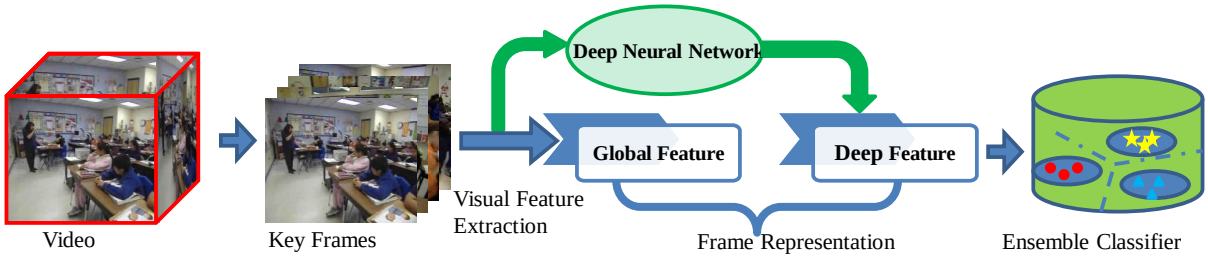

Fig. 2. Pipeline of the proposed educational visual classifier with deep features on online educational videos.

approach for video classification uses the meta-data of videos such as title, speaker's profile, tagged data or abstract associated with the videos. However, when academic videos are crawled and gathered from the Web, identifying and extracting such metadata is non-trivial and erroneous task, as many of videos do not have accurate meta-data to be used. Hence, one crucial aspect to enhance classification performance is to build visual classifiers targeted on educational video frames, which will greatly facilitate users to better index and search video content. A significant body of work has been dedicated to generate visual or multimodal indexes of university style lectures [27][28]. In this context, slides and speech are dominant, and targeted slide and text detection approaches have been shown to work reasonably well. However, there exist a large quantity of educational videos (for example the aforementioned Science shows or from publishers) containing a variety of visual content ranging from natural landscapes to classrooms, from animations to kids, from animals to plants, *etc*. This constitutes a special class of educational videos which thus far has not been investigated under the perspective of visual indexing, since all efforts on educational content indexing have been focused on more traditional lecture style videos.

In order to fill such gap, we propose a set of visual classifiers focusing on categories which are related to educational content, but complement existing efforts on slide detection and OCR.

Inspired by the recent progress of semantic model vector on unconstrained real-world videos [2][17], we investigate the performance of semantic visual classifiers in our application domain containing educational video data. Different from prior work [6][12][5], we explore the effectiveness of deep learning features for video indexing, and employ them as a complementary cue with respect to traditional global descriptors.

In this work, we apply deep learning features obtained from the last layers of a network learned from the ImageNet dataset, to the domain of video indexing. While deep features have been successfully employed for various image classification tasks, their utilization in the video domain has been only recently introduced by Karpathy et al. [7] in the particular domain of action classification. The video indexing domain we explore in this work is made additionally challenging by need for the temporal localization, since a person performing a query is interested only in a given segment of the video, not in the whole clip, especially when the video is long. We address this issue by indexing directly keyframes instead of aggregating scores from keyframes to the entire video clip.

We evaluate our indexing approach on a challenging real-world educational video dataset, which contain $152,919$ Web images as training images and $30,643$ key frames extracted from nonlecture educational videos clips. Table I illustrated the detailed information about our dataset. Figure 2 illustrates the flowchart of our approach. Given an educational video clip, we extract both low level global features and deep features from its key frames. We then employ a set of our pre-built classifiers to predict the concepts appearing in the key frames. The experimental results revealed that the proposed framework employing deep features is able to produce more accurate retrieved results in comparison with global features alone for this educational video search task, achieving a mean Average Precision performance of 0.42.

TABLE I
NUMBER OF KEYFRAMES FROM OUR REAL-WORLD SCALABLE DATASET.

NON-LECTURES	Keyframes	Avg Keyframes/video	Videos
STEMbite	1472	24	61
Scishow	10482	23	450
Crashcourse	17676	80	221
Pearson	867	16	54
NICA	5415	57	95
TOTAL	35912	40	370

The main contributions of this paper are the followings:

- We explore the state-of-the-art deep learning feature extracted with both ImageNet model and self-

trained CNN model. We investigate its performance in comparison with conventional global features.
- We contribute a real-world educational video datasets, including one toy dataset and one scalable dataset. The scalable dataset contain $152,919$ Web images as training images and $30,643$ key frames extracted from nonlecture educational videos clips.
- Targeting on the domain of online educational videos, we conduct extensive experiments with the proposed educational classifier and demonstrate its effectiveness assisted with deep features.

The rest of this paper is organized as follows. Firstly, we provide a review of the related work on video recognition and indexing in Section II. The proposed education visual classifier framework is elaborated in Section III. Specifically, we elaborate global features, low-level features and deep features from Section III-A to Section III-C, respectively. The proposed educational visual classifiers is detailed in Section III-D. In Section IV, we report our experimental results on a real-world educational video dataset, followed by the conclusions in Section V.

II. Related Work

In this section, we provide a briefly description of the existing video indexing approaches, review the deep learning features exploited in recent literature, and describe their application on video domain.

A. Video Indexing

Video recognition research has been largely driven by the advances in image recognition methods, which were often adapted and extended to deal with video data. A large family of video action recognition methods is based on shallow high-dimensional encodings of local spatio-temporal features. For instance, the algorithm of [30] consists in detecting sparse spatio-temporal interest points, which are often described using local spatio-temporal features: Histogram of Oriented Gradients (HOG) [31] and Histogram of Optical Flow [32]. The features are then encoded into the Bag-of-Features (BoF) representation, which is pooled over several spatio-temporal grids (similarly to spatial pyramid pooling) and combined with an SVM classifier. In a later work [33], it was shown that dense sampling of local features outperforms sparse interest points.

Instead of computing local video features over spatio-temporal cuboids, state-of-the-art shallow video representation [35] make use of dense point trajectories. The approach, first introduced in [34], consists in adjusting local descriptor support regions, so that they follow dense trajectories, computed using optical flow. The best performance in the trajectory-based pipeline was achieved by the Motion Boundary Histogram (MBH) [32], which is a gradient-based feature, separately computed on the horizontal and vertical components of optical flow. A combination of several features was shown to further boost the accuracy. Recent improvements of trajectory-based shallow representation include compensation of global (camera) motion [36] and the use of the Fisher vector encoding [37] for trajectory features [35].

B. Deep Learning Features

Deep learning features have been shown to set the state-of-the-art in many applications such as OCR [20], speech recognition [19] and object detection [21]. Researchers have also developed a myriad of approaches as well as toolkits with deep learning features to tackle the challenging problems in various research areas. For instance, Attila [22] and Kaldi [23] for speech recognition, Overfeat [24], CAFFE [3] and Cuda-convnet [9] for visual applications, Theano [25] for natural language processing or more general learning purpose.

There has also a number of attempts to develop a deep architecture for video recognition. In the majority of these works, the input to the network is a stack of consecutive video frames, so the model is expected to implicitly learn correspondences or spatio-temporal motion-dependent features in the first layers, which can be a difficult task. In [38], an HMAX architecture for video recognition was proposed with pre-defined spatio-temporal filters in the first layer. Later, it was combined [39]with a spatial HMAX model, thus forming spatial (ventral-like) and temporal (dorsal-like) recognition streams. Unlike our work, however, the streams were implemented as hand-crafted and rather shallow (3-layer) HMAX models. In [40], a convolutional RBM was used for unsupervised learning of spatio-temporal features, which were then plugged into a ConvNet for action classification. Discriminative end-to-end learning of video ConvNets has been addressed in [41] and, more recently, in [16], who compared several ConvNet architectures for action recognition. Training awas carried out on a very large Sports-1M dataset, comprising 1.1M Youtube videos of sports activities. Interesting, [16] found that a network, operating on individual video frames, performs similarly to the networks, whose input is a stack of frames. This might indicate that the learnt spatio-temporal features do not capture the motion well. The learnt representation, fine-tuned on the UCF-101 dataset, tuned out to be 20% less accurate than hand-crafted stat-of-the-art trajectory-based representation [14].

III. Educational Visual Classifiers

In this section, we illustrate our approach for extracting global features, low-level features and deep features. Then we provide a formal description of the proposed educational visual classifiers on top of them.

A. Global Features

We extract three types of global visual descriptors including color, texture and edges to represent each image.

- **COLOR**: Color descriptors consist of color correlogram, color histogram and color wavelet features. In HSV color space, Color correlogram is a global color and structure feature which can be represented by 166-dimensional single-banded auto-correlogram using 8 radii depths. Color histogram is a 166-dimensional histogram feature vector in the same HSV space. Color moments are extracted from a 5*5 grid and represented by the first 3 moments for each grid region in Lab color space as a normalized 255-dimensional vector.
- **TEXTURE**: Texture descriptor contains wavelet texture, GIST and LBP histogram features. Wavelet texture is localized texture extracted from a 3*3 grid and represented by the normalized 108-dimensional vector of the variances in 12 Haar wavelet sub-bands for each grid region. To compute the GIST descriptor, the image is segmented by a 4 by 4 grid for which 8 orientation histograms of 4 scales are extracted and produces a 512-dimensional feature vector. The LBP histogram feature contains 58 uniform and 1 non-uniform pattern.
- **EDGE**: the Edge histogram feature is computed with 8 edge direction bins and 8 edge magnitude bins which is based on a Sobel filter, resulting in a 64-dimensional feature vector.

We observe that having a large diversity of visual descriptors is important for capturing different semantics and dynamics in the video scene, as reported in previous works [2]. Moreover, we extract the descriptors at different spatial granularities (i.e., global, layout, pyramid, horizontal parts, grid, horizontal center, vertical center). Such spatial divisions have shown improved performance and robustness in video retrieval benchmarks [1]. In sum, we extracted 47 different kinds of features for the classifiers training.

B. Low-level Features with Fisher Vector Representation

Fisher Vector (FV) coding approach, derived from Fisher Kernel, was originally proposed for large scale image classification. Compared with other coding methods such as vector quantization and sparse coding, FV coding can easily obtain high-dimensional feature codes with small codebook size, which has been shown to provide considerable performance improvements when utilizing linear classifiers.

Following best practices reported in the literature [14], we process the descriptors independently. We extract SIFT low-level visual features and leverage the GMM model followed by Fisher Vector over the features extracted from all the Web images. Specifically, we first densely extract local SIFT descriptors with a spatial stride of 4 pixels at 9 scales and the width of SIFT spatial bins is fixed as 8 pixels, which are the default settings in the VLFeat toolbox [42]. We learn a GMM dictionary sampled from a subset of one million SIFT descriptors. All descriptors are whitened after PCA processing to 64-dimension with a ratio of 0.5. We then conduct FV encoding and apply ℓ_2 normalization to the resulting super vectors.

C. Deep Features

1) ImageNet Model: Inspired by the success of deep learning framework on large-scale visual recognition tasks such as classification, embedding and object detection [21], we first employ it as the feature extractor for video frames. For the implementation, we utilize the open source deep learning framework CAFFE [3], which is based on the deep convolutional neural network architecture by Krizhevsky et al. [9]. In fact, our testing query and database frames are independent from the ImageNet dataset, hence we use the pre-trained ImageNet model for ILSVRC image classification from [18], as an analog to using the prior knowledge a human obtained from previous visual experiences to learn new tasks more efficiently.

The activations of the neurons in the late hidden layers could be used as strong features for a variety of video recognition tasks because they contain much richer semantic representations than any earlier convolutional layer in the network. Therefore, we use features from the last three layers in our implementation, which are the first set of activations that have been fully propagated through the convolutional layers of the network, the final hidden layer (i.e., just before propagation through the final fully connected layer to produce the class predications) and the layer between them. We set the network input with mean-centered raw RGB pixel 256 $\times$ 256 images, the values are forward propagated through 5 convolutional layers (i.e., pooling and ReLU non-linearities) and 3 fully-connected layers (i.e., to determine its final neuron activities), finally we obtain the

4096-dimension vector for the last hidden layers and normalize the vectors by their ℓ_2 normalization.

2) Our CNN Model: We train our own model to obtain the deep feature representation based on the Convolutional Neural Network (CNN). The convolutional network consists of several layers and each layer is a linear transformation followed by a non-linear one. The first layers takes an $227 \times 227 \times 3$ input image as the input. The network is based on the architecture used in [3]. Each layer consists of: (1) convolutional of the previous layer output with a set of filters; (2) passing the responses through a rectified linear function. To obtain the CNN model, we divide our dataset (including $152,919$ images) into two parts,the training part contains 80% of Web images and validation dataset with the rest of 20% images. We describe the in

TABLE II
GLOBAL VISUAL FEATURES EXTRACTED FOR VISUAL CLASSIFIERS.

Color	Texture	Edge
color correlogram	lbp histogram	edge histogram
color histogram	GIST	-
color wavelet	wavelet texture	-

TABLE III
NOTATIONS AND DESCRIPTIONS.

Notations	Descriptions
m	chi square, histogram intersection, linear approximate
S_b	maximum num of positive samples per unit model
N_b	maximum num of bags to try for each descriptor
r_d	fraction for unit model training
w	sum or weighted sum
N_p	num of parameters to search for modeling

D. Educational visual model learning

For each feature type, we learn classifier models from a number of N_b of bags of training data, randomly sampled with a majority of positive and negative samples, with sample ratio r_d. Specifically, one-versus-all SVMs with various kernels (Chi-square, Histogram Intersection, Linear Approximate Kernel) are trained, independently for each concept, based on each descriptor. During training for one category, all the samples from the other categories are used as the negative examples. The default parameters for N_b and r_d used to train all the base models for the semantic models are 3 and 0.3, respectively, which result in a pool of $N = 141$ base models for each concept. To minimize the sensitivity of the parameters for each base model, we choose the SVM parameters based on a grid search strategy. Then we build the SVM models with different values on the kernel parameters C, the relative cost factors of positive versus negative samples, the feature normalization schemes, and the weights between training error and margin. Each model is then associated with its cross validation performance, where average precision is employed as the performance measure. Moreover, the fusion strategies of the base models into an ensemble classifier are determined based on average precision performance on the held-out data. To reduce the risk of overfitting, we control the strength and correlation of the selected base models by employing a greedy model selection step. The algorithm iteratively selects the most effective base model from the unit models pool, adds it to the composite classifier without replacement, and evaluates its average precision on the held-out set. Once the best parameters are determined, the SVMs are retrained on the whole development set. The semantic model output is then the ensemble classifier with the highest average precision observed on the held-out set. For clarity, we illustrate several important notations and their definitions throughout the experiments in Table III.

IV. EXPERIMENTS

To enhance the educational classifier performance, we first conduct extensive experimental comparisons tuned with sets of various parameters. Then, we investigate deep feature performance in comparison with conventional global features.

The experiments are performed in two datasets: a toy datset and a real-world scalable dataset. The toy dataset is mainly designed for parameter tuning while the other dataset is utilized the demonstrate the effectiveness of the proposed educational classifiers.

We implement the experiments on a server with a 8-core, 2.4GHz and 24 GB memory. We adopt Average Precision, a common choice for evaluating visual search engines [1][2],

$$\mathrm{AP} = \frac{1}{TP} \sum\nolimits_{d=1}^{N} \frac{TP_d}{d} I_d \tag{1}$$

where $I_d = 1$ if the d^{th} sample is relevant and 0 otherwise. TP is the total number of relevant samples in the collection and N is the number of total samples. TP_d denotes the number of relevant samples found in the top d ranked samples returned by the system. In addition to Average Precision, we also report overall performance of mean Average Precision (mAP) which is averaged over the concepts.

A. Data and Methodologies

Pearson Toy Dataset: We first collect 22,037 Web images from 16 popular concepts via public api as the toy

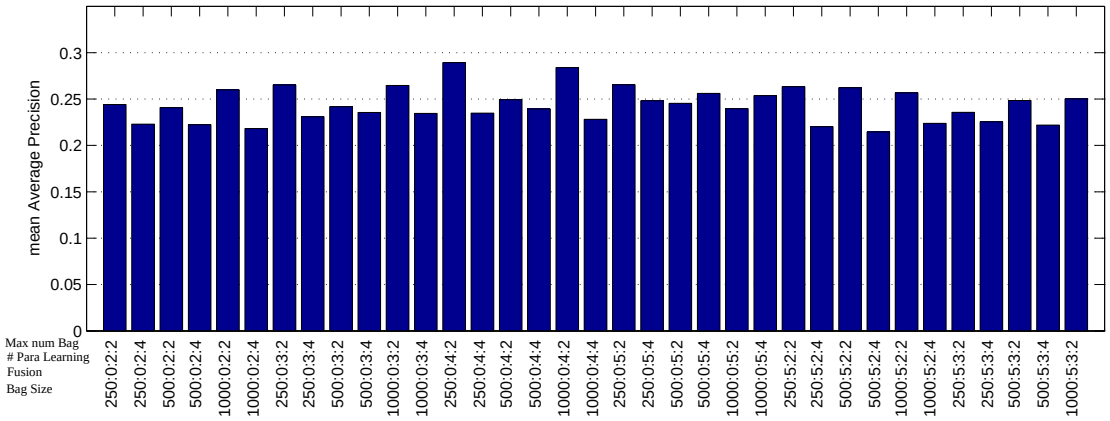

Fig. 4. Retrieval performance comparison in mAP with global features in out dataset. The results are tuned with variables (a) max number of bags (b) number of parameter learning (c) fusion approach(sum or weighted sum) (d) bag size.

Fig. 3. Sample images from the training dataset.

training dataset: “Boy”, “CGI”, “Child”, “Classroom”, “Dog”, “Flower scene”, “Girl”, “Greenery”, “Hands”, “Head and shoulders”, “Human face”, “Kitchen”, “Mountains”, “Printed text”, “Puppet”, “Reading”. The test dataset contains 54 clips of videos with average 120 seconds duration provided from Pearson education collection. We extract and select 1005 key frames from these video clips. For each frame, its relevance to the corresponding query is labeled with two levels: relevant and irrelevant which indicated by scores 1 and 0, respectively. The ground truth of each frame is manually labeled by professional image researchers. Some exemplar images of each concept are illustrated in Figure 3 and number of images and extracted video frames in training and test dataset are detailed in Table IV.

Scalable Dataset: In order to further validate the proposed educational classifiers, we extend the toy dataset in a large-scale manner. The scalable dataset contains $152,919$ Web images. We crawled and organized the training dataset in a hierarchical structure which contains 6 main classes and 162 subcategories. The main classes include “Animals”,“People”, “objects”,“Setting”, “Graphical Element” and “Activities”. Table **??** illustrated the detailed information about the dataset. The keyframes in the test dataset includes $35,912$ samples from 370 video clips.

B. Performance Evaluation in Pearson Toy Dataset

We perform experimental evaluation with one-versus-all classifiers using global features. During training for one concept, all the images from the other concept are used as negative examples. In order to optimize the sensitivity of the parameters for each base model, we select parameter C from $\{0.1,1,10,100\}$ via grid search on a 5-fold cross validation, with a 70% training and 30% validation random splits on both positive and negative examples of datasets. To evaluate the retrieval performance, we empirically choose and tune four sets of important variables consisting of max number of bags, number of parameter to search for modeling, fusion method and bag size. Figure 4 provides the detailed performance comparison with various parameter combinations in terms of mAP. It obtains 0.289 in mAP score when we use 250 samples per bag and 2 bags per unit model in the training set. We observe that it achieves better performance when we set bag size as two in comparison with four bags. This is because the over-fitting problem might occur in the experiment if we adopt higher number bags per feature.

Also, we conduct an experimental comparison between classifier with linear approximate kernel [29] and

TABLE IV
NUMBER OF IMAGES AND EXTRACTED VIDEO FRAMES IN PEARSON DATASET.

Concepts	Training	Test	Concepts	Training	Test
Boy	544	101	Greenery	1068	15
Reading	1217	145	Classroom	804	31
CGI	438	99	Head&Shoulders	1625	340
Child	647	191	Human face	4906	220
Puppet	214	128	Kitchen	564	10
Dog	2352	14	Mountains	843	10
Hands	1200	73	Flower scene	825	10
Girl	710	86	Printed text	4080	342

non-linear kernel (Chi-square, Histogram-intersection). When fixing the other parameters, we observe that the mAP of linear approximate kernel achieves 0.23 while non-linear kernel yields an mAP of 0.28. Hence, non-linear kernel performs favorably and better than linear approximate kernel in this search task.

TABLE V
PERFORMANCE COMPARISON BETWEEN DEEP FEATURES AND GLOBAL FEATURES IN MAP.

Feature Type	mAP	improvement
[29]	0.23	-
Global Feature	0.28	+21.7%
Deep Feature	**0.42**	+82.0%

To enhance retrieval performance of educational classifier, we conduct the experimental comparison with deep learning features. We use the same parameter setting as selected in the previous subsection. The mean Average Precision overall all the concepts based on our algorithm is also shown in Table V. Obviously from the observation above, our method with deep features achieves a mAP of 0.42 and obtains a noticeable improvement compared with global features. The performance gives credits to the deep features by enhancing the discriminative power of educational classifier. Hence, assisted with deep learning features, the experimental result demonstrates superiority of our approach over conventional global features. Nevertheless, we notice that our algorithm is less effective in the concepts "Puppet" and "Human face". And this may result from that they share a couple of images which belong to both two categories.

To further test the effectiveness of our proposed approach, we validate our performance in comparison with the LibLinear classifier [11] with deep features. Figure 7 provides the experimental comparison in Average Precision among all concepts and presents the score comparison in terms of mAP. It can be observed that our approach achieved significant improvement when mAP score increased from 0.21 to 0.42. The above

Feature	mAP
SIFT+FV	53.16%
GLOBAL FEATURES	33.97%
Deep Feature with Pre-trained Model	61.55%
Self-trained Deep Model	61.55%

TABLE VI
PERFORMANCE COMPARISON OF HETEROGENEOUS FEATURES ON EDUCATIONAL VIDEO INDEXING.

results demonstrate that the proposed approach with deep feature is more suitable in the task of educational video search.

C. Large-Scale Evaluation with Self-Trained Model

In order to train the deep model, we divide the scalable dataset into two parts. Specifically, the training part contains 80% of Web images and validation dataset with the rest of 20% images. We implement the network under the caffe framework [3] and train them using stochastic gradient descent with a batch size of 256 examples, momentum of 0.9, and a weight decay of 0.0005. We use an equal learning rate for all layers and adjust it every 10,000 iterations. The strategy is to divide the learning rate by 10 when the validation error rate stops decreasing with the current learning rate. The learning rate is initialized to 0.01 for the network. We set the maximus iterations as 50,000. We train the networks for roughly 65 epochs, which takes 28 hours on one NVIDIA K40c GPU. It can rapidly process one 227 * 227 image within about 0.5 second.

V. CONCLUSION

In this paper, we investigated the state-of-the-art deep learning feature and evaluated its performance compared with conventional global features in the domain of educational video content retrieval. We introduced an educational video retrieval system based visual classifiers assisted with deep features, and evaluated it by conducting retrieval experiments on a real-world dataset of online learning videos. The extensive experiments demonstrated the effectiveness of our proposed framework in this educational video retrieval task.

REFERENCES

[1] M. Campbell, A. Haubold, M. Liu, A. Natsev and J. R. Smith. IBM research TRECVID-2007 video retrieval system. In *NIST TRECVID Workshop*, 2007.
[2] M. Merler, B. Huang, L. Xie, G. Hua and A. Natsev. Semantic model vectors for complex video event recognition. In *IEEE Transaction on Multimedia*, vol. 14, no. 1, pp. 88-101, 2012.
[3] Y. Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. *http://caffe.berkeleyvision.org/*, 2013.
[4] I. Allen and J. Seaman. Changing course: ten years of tracking online education in the united states. In *Annual Report of Babson Survey Research Group*, 2013.

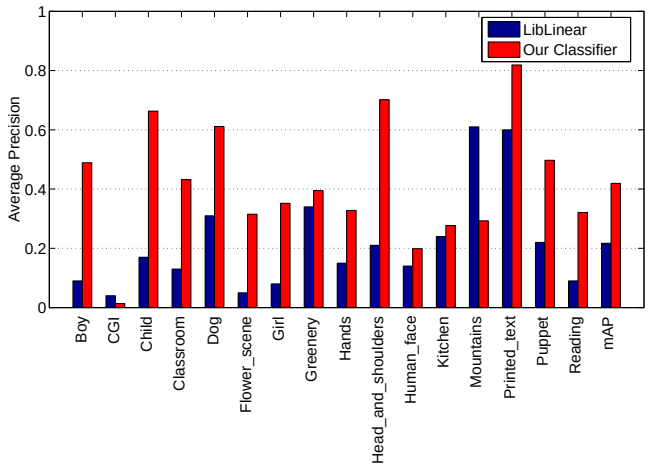

Fig. 5. Detailed performance comparison between Liblinear classifier and educational ensemble classifier.

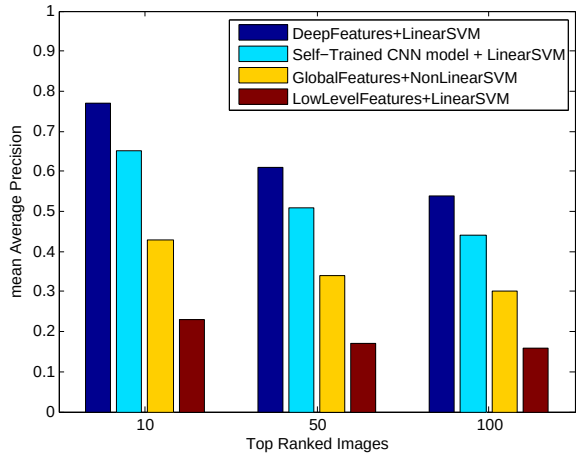

Fig. 6. Performance comparison in mean Average Precision (mAP) with low-level features, global features and deep features.

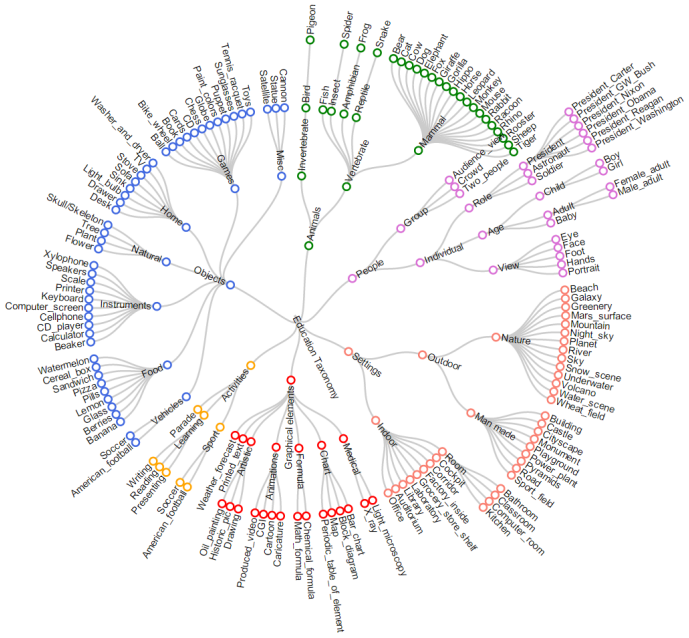

Fig. 8. Visualization of hierarchical taxonomy tree for 162 categories in the training dataset.

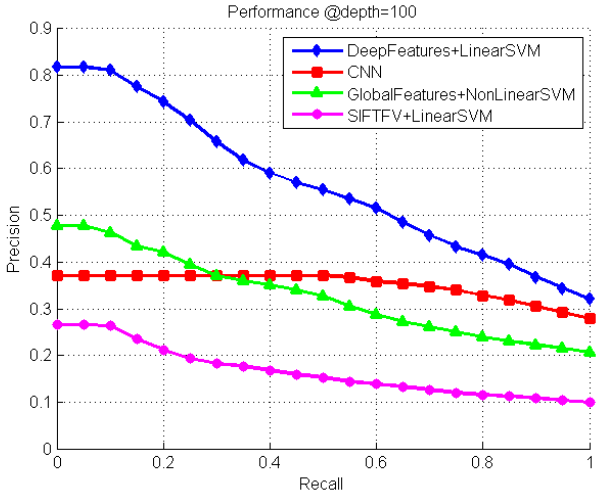

Fig. 9. Performance comparison in precision and recall at top 100 ranked images.

[5] R. Yan, T. Tesic and J. R. Smith. Model-shared subspace boosting for multi-label classification. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2007.
[6] J. Smith, M. Naphade and A. Natsev. Multimedia semantic indexing using model vectors. In *Proceedings of the IEEE Conference on Multimedia and Expo*, 2003.
[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.
[8] M. Zeiler and R. Fergus. Visualizing and understanding convenlutional networks. In *Proceedings of the IEEE European Conference on Computer Vision*, 2014.
[9] A. Krizhevsky, I. Sutskever and G.E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advanced in Neural Inforamtion Processing System*, 2012.
[10] G. Griffin, A. Holub and P. Perona. The caltech 256. In *Caltech Technical Report*, 2006.
[11] R. Fan, K. Chang, C. Hsieh, X. Wang and C. Lin. LIBLINEAR: a library for large linear classification. In *Journal of Machine Learning Research*, vol. 9, no. 4, pp. 1871-1874, 2008.
[12] S. Ebadollahi, L. Xie, S. Chang and J. Smith. Visual event detection using multi-dimesional concept dynamics. In *Proceedings of the ACM International Conference on Multimedia and Expo*, 2006.
[13] A. Natsev, M. Naphade and J. Smith. Semantic representation, search and mining of multimedia content. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004.
[14] X. Peng, L. Wang, X. Wang and Y. Qiao. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. In *arXiv*, 2014.
[15] A. Razavian, H. Azizpour, J. Sullivan and S. Carlsson. CNN Features off-the-shelf: an astounding Baseline for recognition. In *arXiv*, 2014.
[16] K. Simonyan and A. Zisserman. Two-stream conventional networks for action recognition in videos. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014.
[17] X. Li, C. Snoek, M. Worring, D. Koelma and A. Smeulders. Boostrapping visual categorization with relevant negatives. In *IEEE Transaction on Multimedia*, vol. 15, no. 4, pp. 933 - 945, 2013.
[18] Large scale visual recogntion challenge. *http://www.image-net.org/challenges/LSVRC/2012/*, 2012.
[19] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed and N. Jaitly. Deep neural networks for acoustic modeling in speech recognition: The shared

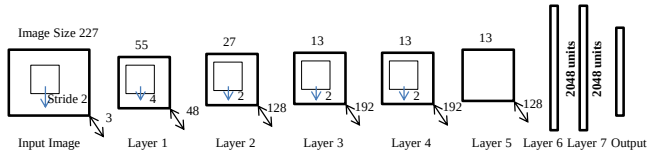

Fig. 10. Architecture of our CNN network. We take a 227 × 227 × 3 image as the input and convolve it with 48 different first layer filters, each of which with the size 7 × 7, using a stride of two in both dimensions.

views of four research groups. In *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2008.

[20] Y. Lecun, K. Kavukcuoglu and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2010.

[21] O. Russakovsky, J. Deng, H. Su and A. Berg. ImageNet large scale visual recogntion challenge. In *arXiv*, 2014.

[22] H. Soltau, G. Saon, and B. Kingsbury. The IBM Attila speech recognition toolkit. In *Proceedings of the IEEE Spoken Language Technology Workshop*, 2010.

[23] D. Povey et al. Deep Neural Networks in Kaldi. *http://kaldi.sourceforge.net/dnn.html*, 2013.

[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In *arXiv*, 2014.

[25] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. Theano: A CPU and GPU Math Expression Compiler. In *Proceedings of the Python for Scientific Computing Conference*, 2010.

[26] A. Bergamo and L.Torresani. Classemes and Other Classifier-based Features for Efficient Object Categorization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.36, no.10, pp.1988 - 2001, 2014.

[27] A. Haubold and J. Kender. VAST MM: Multimedia Browser for Presentation Video. In *Proceedings of the ACM International Conference on Image and video retrieval*, 2007.

[28] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash and L. Rowe. TalkMiner: A lecture webcast search engine. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[29] L. Cao et al. IBM Research TRECVID-2012 Multimedia Event Detection (MED) Systems. In *Techical Report*, 2012.

[30] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[31] N. Dalal and B. Triggs. Histogram of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[32] N. Dalal and B. Triggs. Human detection using oriented histograms of flow and appearance. In *Proceedings of the IEEE European Conference on Computer Vision*, 2006.

[33] H. Wang, M. M. Ullah, A. Klaser, I. Laptev and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the IEEE British Machine Vision Conference*, 2009.

[34] H. Wang, A. Klaser, C. Schmid and C. L. Liu. Action recognition by dense trajectories. In *Proceedings of the IEEE European Conference on Computer Vision*, 2011.

[35] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[36] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recogntion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[37] F. Perronnin, J. Sanchez and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proceedings of the IEEE European Conference on Computer Vision*, 2010.

[38] H. Jhuang, T. Serre, L. Wolf and T. Poggio. A biologically inspired system for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[39] H.Kuehne, H. Jhuang, T. Poggio and T. Serre. HMDB: A large video database for human motion recogntion. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[40] G. W. Tayleor, R. Fergus, Y. LeCun and C. Bregler. Conveolutional learning of spatio-temporla features. In *Proceedings of the IEEE European Conference on Computer Vision*, 2010.

[41] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.35, no.1, pp.221 - 231, 2013.

[42] A. Vedaldi and B. Fulkeson. VLFeat: An open and portable library of computer vision algorithms. 2008.

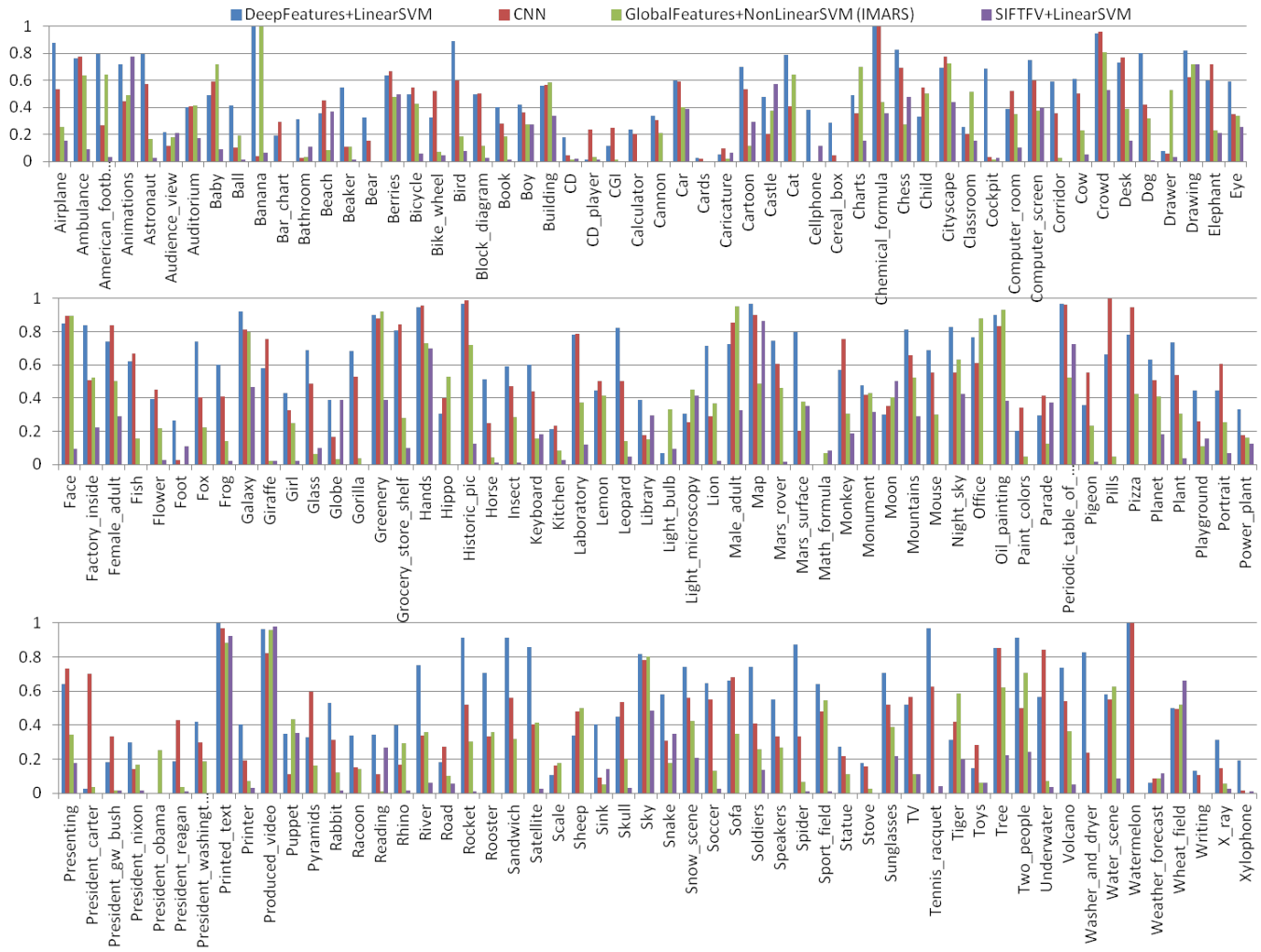

Fig. 7. Detailed experimental comparison of 162 categories Average Precision (mAP) with low-level features, global features and deep features.