Modality Classification for Searching Figures in Biomedical Literature

Zhiyun Xue[1], Md Mahmudur Rahman[2], Sameer Antani[1], L. Rodney Long[1], Dina Demner-Fushman[1], George R. Thoma[1]

[1]Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, USA
{xuez, santani, rlong, ddemner, gthoma}@mail.nih.gov

[2]Computer Science Department
Morgan State University
Baltimore, USA
md.rahman@morgan.edu

***Abstract*-—Image modality classification categorizes images according to their type. It is an important module in the Open-i^{SM} multimodal (text+image) search engine that retrieves figures from biomedical articles. It is a hierarchical classification where on the top level the input figures are classified into two general categories: regular images (X-ray, CT, MRI, photographs, etc.) vs. illustration images (cartoon sketch, charts, graphs, etc.). This binary classification task is challenged by the vast diversity of visual material (image type), and the way it is organized (simple or compound figures). We present two methods for this binary classification: (i) Support Vector Machines (SVM) with manually-selected features, including a feature based on semantic concepts; and, (ii) Deep Learning method which avoids the process of feature handcrafting. Both methods were tested and compared on a dataset of 16400 figures. Both methods achieved good performance (above 95% accuracy). The slightly better performance of the feature-based method demonstrates the effectiveness of the features we chose.**

Keywords—Modality classification; figure searching; concept feature; support vector machine; deep learning; convolutional neural networks

I. INTRODUCTION

The classification of images based on their visual type (modality) is an important step in medical image retrieval systems. Modality image classification provides an option to limit the search space that users are interested in, and also improves the retrieval performance of the system. Modality classification has been integrated into Open-i^{SM} [1], a multimodal (image + text) search engine for biomedical literature that has been developed by the U.S. National Library of Medicine (NLM). Unlike the images in other medical image retrieval systems, which may operate, for example on clinical images from PACS systems, the figures that appear in the biomedical literature are much more diverse and contain many non-medical images. Figure 1 shows the hierarchy of the modality classification used by ImageCLEF 2015 [2] (ImageCLEF is an annual competition aiming to provide an evaluation forum for image annotation and retrieval). The first level of the hierarchy is *diagnostic images* vs. *illustrations*. The category of *diagnostic images* contains images in the categories of *radiology*, *printed signals*, *microscopy*, *visible light photos* and *3D reconstructions*. The *illustrations* include images of *tables*, *forms*, *charts*, *gene sequences*, and so forth. Regarding this first level of modality classification, Open-i uses a slightly different arrangement. In Open-i, the *gel chromatography* and *non-clinical photos* subcategories inside the *illustrations* category are moved to the *diagnostic images* category and the *diagnostic images* category is renamed as *regular images*, based on the observation that the visual characteristics of images in those two subcategories are more similar to those of *diagnostic images* than to those of other illustrations. Similarly, the subcategory of *printed signals* (*waves*) is moved to the *illustrations*. In this paper, we will present our work on classifying figures in two categories: *regular* vs. *illustration*, which is the first level of the modality classification hierarchy used in Open-i.

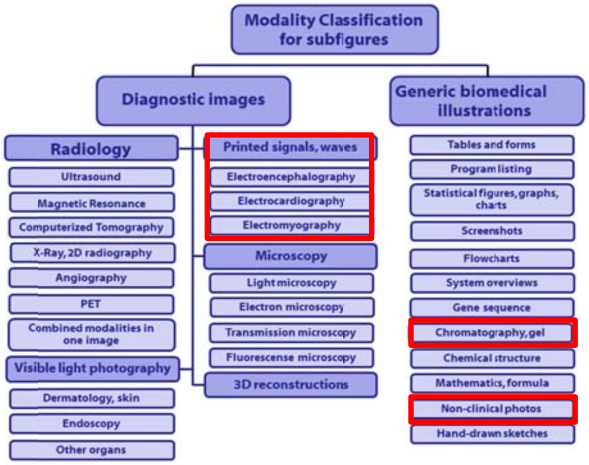
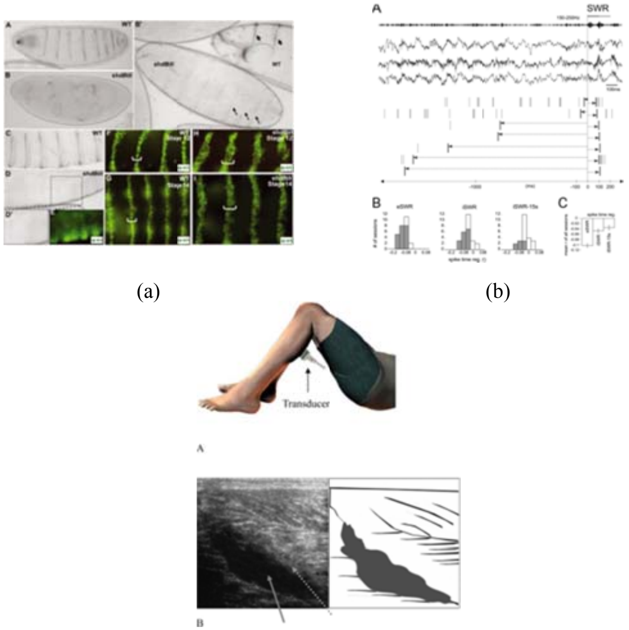


Fig. 1. Modality classification hierarchy used in ImageCLEF 2015

The modalities defined in Figure 1 are for single-panel figures or subfigures. Very often, the figures in biomedical literature are *compound* figures (also called *multipanel* figures). A compound figure may consist of subfigures that are of different modalities in different levels. For example, Figure 2 exemplifies three cases. All the subfigures in Figure 2(a) are *regular* images (may be in different subcategories of *regular* images). All the subfigures in Figure 2(b) are *illustration* images (may be in different subcategories of *illustration* images). The subfigures in Figure 2(c) are a combination of *regular* images and *illustration* images. We have developed a method for splitting the multipanel figures and matching the available labels (such as *A*, *B*, *C*, …) to each of the separated subfigures [3]. The method is very effective for images in the *regular* category, and our results for the compound figure separation task in ImageCLEF 2015 are the best among all the participants [4]. However, the method does not produce sufficiently satisfactory results for *illustration* compound figures to allow the subfigures to be integrated into Open-i. At present, the step of regular/illustration classification is put before the step of figure panel splitting. Therefore, the dataset we create to test the proposed methods contains not only single-panel figures but also compound figures. As a result, the challenges we face include not only the vast diversity of visual content presented in the single panel figures but also different combinations of subfigures in different subcategories.



(a) (b)

(c)

Fig. 2. Multipanel figures in which panels belong to different modalities

In this paper, we present two approaches for classifying figures into two general categories: *regular* and *illustration*. One is based on the traditional approach which contains the step of extracting specific features. The other is based on a deep learning approach, which avoids the process of feature handcrafting and can automatically discover multiple levels of representations from raw input data. We compared these two methods on a dataset consisting of 16400 figures. Both methods achieved good performance levels (above 95%). The performance of the traditional approach is slightly better (3% higher), which indicates the high effectiveness of our features for this application, especially one feature that our group developed for reducing the semantic gap (a problem low-level visual features usually suffer from).

The rest of the paper is organized as follows. Section II describes the two methods. Section 3 presents the experimental testing results and comparisons. The conclusions and future work are presented in Section 4.

II. Method

A. Feature based classification

The conventional approach generally consists of two main components: features and classifier. The goal of the feature extraction step is to obtain a set of features that represent the visual characteristics of the original data and capture the perceptional characteristics that discriminate among the images in different categories. The effectiveness of the features is dependent on the specific application at hand. Therefore, it often requires domain knowledge and engineering skills, as well as experimental trial and error to find a good set of features. Given the feature vectors calculated from the original images and the corresponding labels of the images, supervised classification algorithms can learn to associate ground-truth labels with these training data, and then may be used to predict labels for images whose truth labels are unknown. A range of such algorithms is available, including decision trees, support vector machines (SVM), random forest, and neural networks, each having characteristic strengths and weaknesses. In the following, we will introduce the features and classifier we used for this specific classification task. Based on the experimental results (reported in Section 3), the features and classifier we employed were very effective.

1) Features

We applied several types of features, including the semantic concept feature we developed.

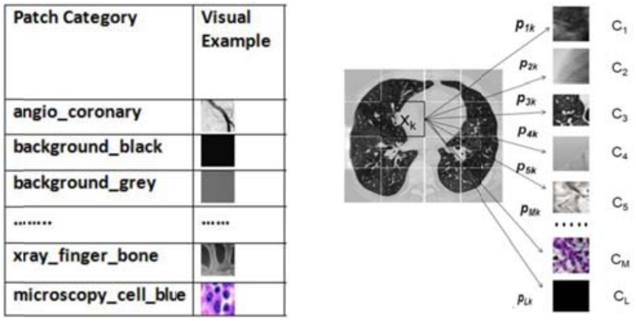
- Semantic concept feature

A major component of any image-based classification system is the feature representation of images in terms of low-level features (color, texture, edge, shape, etc.) or the recently popular "Bag of Visual Words" (BoVW) features [5]. In the BoVW approach, generally the low-level visual features of local regions of points, such as color, texture, and so forth, are vector quantized to generate the visual words. Although it has proved to be effective for image representation (and is similarly effective for document representation in text retrieval), the unsupervised clustering to generate the words or dictionary largely neglects the semantic contexts of the local features. As a result, commonly generated visual words are still not as expressive as keywords in text documents.

In a heterogeneous collection of medical images of journal articles, it is possible to identify specific local patches that are perceptually and/or semantically distinguishable, such as homogeneous texture patterns in grey level radiological images, or differential color and texture structures in microscopic pathology and dermoscopic images. For example, if we consider a computed tomography (CT) image of the lung or chest, which appears in many radiographic or medical journals, we observe several image regions with texturally different visual patterns that are semantically distinguishable from each other, such as a slightly bright and hazy appearance, can be mapped to the pattern "Ground Glass" opacity, or hexagonal structures that can be mapped to a "Honey

Combing" pattern. The variation in these local patches can be effectively modeled by using supervised classification techniques. A supervised learner can create a model of different visually interesting patterns as concepts to capture the variability of the local patches with sufficient training samples. In this context, an instance (e.g., local patch) in the training set can be represented by a feature vector along with its local concept or category-specific label [6].

Therefore, in order to perform concept-based image representation (e.g. "Bag of Concepts"), we at first manually annotated a set of training concepts from distinguished local image patches to perform supervised learning, as done by a method such as SVM [7]. Our goal is to learn a set C of L labels, where C = {c1,··· ,ci,··· ,cL}, and where each ci ∈ C characterizes a visual concept. The training set of the local patches are generated by a fixed-partition based approach and represented by a combination of color and texture moment and edge frequency related features.



Patch Category	Visual Example
angio_coronary	
background_black	
background_grey	
....	......
xray_finger_bone	
microscopy_cell_blue	

Fig 3. Example image patches (left) and concept ***(C₁-C_L)*** based image annotation process (right)

We used a multi-class SVM-based classification method which combines all possible pairwise comparisons of binary SVM classifiers, known as one-against-one or pairwise coupling (PWC) [8]. For SVM training, the initial input to the system was the feature vector set of the patches along with their corresponding (manually assigned) concept labels. Images in the data set were annotated with local concept labels by partitioning each image I_j into an equivalent r × r grid of sub-images, where each sub-image represents a *d*-dimensional combined feature vector of color and texture moments. For each sub-image (region), the local concept category probabilities are determined by the prediction of the multi-class SVMs, and finally the category label of the region is determined as the category with the maximum probability score. Hence, the entire image is represented as a two-dimensional index linked to the concept labels as shown in Fig. 3. Based on this encoding scheme, an image I_j can be represented as a vector in a local semantic concept space as

$$\mathbf{f}_j^{\text{Concept}} = [f_{1j},\cdots,fi_j,\cdots fL_j]^{\text{T}} \qquad (1)$$

where each f_{ij} corresponds to the normalized frequency of occurrence of a concept $\boldsymbol{c_i}$, $\mathbf{1 \leq i \leq L}$ in image I_j.

- Other features

Besides the semantic concept feature, we also applied four additional features that have been used in image retrieval. They are: CEDD (color and edge directivity descriptor), FCTH (fuzzy color and texture histogram), CLD (color layout descriptor), and EHD (edge histogram descriptor). CEDD [9] and FCTH [10] are two descriptors used by the Lucene image retrieval (LIRE) library for image indexing and retrieval. Both features incorporate color and texture information in one histogram which results from the combination of three fuzzy units. The first and second fuzzy units, the part for color information representation, are the same for CEDD and FCTH. They differ in the third fuzzy unit which is for the capture of texture information. Both features are compact and their sizes are limited to less than 72 bytes per image. CLD and EHD are MPEG-7 features [11]. CLD captures the spatial layout of the dominant colors on an image grid consisting of 8 by 8 blocks and is represented using DCT (discrete cosine transform) coefficients. EHD represents the local edge distribution in the image, i.e., the relative frequency of occurrence of five types of edges (vertical, horizontal, 45-degree diagonal, 135-degree diagonal, and non-directional) in the sub-images.

2) Classifier

For the classifier to classify the image, we also use the support vector machine (SVM). For the SVM used by the semantic concept feature for classifying local patches, we used the LibSVM [12] Java package for SVM training and testing. Specifically, we use the *C*-support vector classification (*C*-SVC) [12] SVM formulation. For the SVM used for classifying the image, we used the Sequential Minimal Optimization (SMO) algorithm implemented in the Weka [13].

B. Convolutional neural networks

Unlike conventional methods which are based on handcrafted feature extractors, deep learning seeks to automatically obtain good features through a learning procedure. Deep learning methods employ architecture with multiple layers, in which the deeper the layer, the more abstract representation the learning yields. The convolutional neural network (CNN) is one particular type of deep neural network [14]. It was originally proposed in the 1990s [15]. It reignited a lot of interest after it achieved the best result in the ImageNet [16] competition in 2012. Since then, it has become the leading deep learning method used for image classification and object recognition. Therefore, we applied it to our application.

CNNs explore spatial relationships of pixels in images to reduce the number of parameters in the neural network that must be trained. There are four important ideas used by CNNs: local connections, shared weights, pooling and many layers [14]. A typical CNN architecture consists of a number of convolutional and subsampling layers followed by several fully connected layers. The convolutional layer contains several feature maps. Each unit in each of the feature maps is connected to a local subset of units in the feature maps of the previous layer. Mathematically speaking, each feature map is obtained by convolving the input with a linear filter, adding a bias, and then passing through a non-linear function. The subsampling layer usually computes the maximal value of a local subset of units in each feature map in the convolutional layer. This process not only reduces the computational complexity for subsequent layers, but also provides a certain degree of shift-invariance. The fully connected layers are traditional multilayer perceptron (MLP). The parameters of CNNs (weights and biases) are trained by using the back propagation algorithm.

For our application, we use the open-source implementation named *cuda-convnet* [17] which uses Graphical Processing Units (GPUs) to accelerate the computation speed. It was developed by Krizhevsky et al. [18]. *cuda-convnet* provides a number of options, including various types of layers, and hidden unit nonlinearities. There are some one channel grayscale images in the dataset so we convert those one channel grayscale images to three-channel images, and then resize all the images to 32 x 32. Figure 4 shows the architecture of the CNN we apply to our dataset. It contains two convolutional layers (*conv1* and *conv2*), two pooling layers (*pool1* and *pool2*), two locally-connected layers with unshared weights (*local1* and *local2*), a fully-connected layer (*fc*), and a soft max layer (*softmax*). For all the layers except the *fc* layer and the *softmax* layer, we employ rectified linear units (ReLUs) as the nonlinear function.

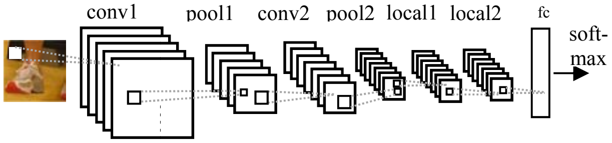


Fig. 4. CNN architecture used

III. Experimental Test

A. Data

The images used in the experiment were downloaded from Open-i. They are figures from articles in the PubMed [19] database. The images are labeled as either regular images or illustration images by visual examination. For a compound figure, if all the subfigures are regular (or illustration) images, then it is labeled as regular (or illustration). The dataset does not contain any compound figure which is composed of subfigures in both categories. The final dataset consists of 8200 regular figures and 8200 illustration figures. The testing set is made up of 2200 regular images and 2200 illustration images randomly selected from the dataset. The training set consists of the rest of the figures, which contains 6000 regular images and 6000 illustration images.

B. Model for semantic concept feature

For concept model generation based on the SVM learning, 60 local concept categories were manually defined from image patches. The local concepts were selected to reflect meanings useful to physicians because of distinct visual appearances, such as different lung tissue patterns of X-ray and CT images, microscopic images of different color and texture patterns, and so on, as shown in Fig. 3. The training set used for this purpose was created by an engineer and consists of around 19,000 patches to represent the 60 concept categories. To generate the local patches, each image in the training set was re-sized to 256 x 256 pixels and partitioned into an 8 x 8 grid generating 64 non-overlapping regions of size 32 x 32 pixels. Only the regions that conform to at least 80% of a particular concept category were selected and labeled with the corresponding category label. Color moments, Auto-Correlation and Edge frequency-based features were extracted and combined to form a 59-dimensional feature vector for all training patches.

For the SVM training (for local patch concepts), we utilized both the radial basis function (RBF) and the polynomial kernels. There are two tunable parameters for RBF kernels: C and γ, and the best values for C and γ cannot be known a priori, for a particular classification task. We adopted the standard solution of using a 10-fold cross-validation (CV) to select these values, where we let C and γ vary over a range of plausible values. (Basically, pairs of (C, γ) were used and the one with the best CV accuracy was picked.) We also experimented with the polynomial kernel of degree 1 and 2 with $C = 100$. However, the best accuracies were achieved by using the RBF kernel as shown in Table 1. After finding the best values of parameters C and γ for the RBF kernel, they were used for the final training to generate the model file for the concept learning.

Table 1. CV accuracies of local concept classification (SVM)

Kernel	C	γ	Degree	Accuracy
RBF	100	0.08		76.03 %
Poly	100		1	74.65 %
Poly	100		2	74.09 %

C. Image classification result

Table 2 lists the classification accuracies for using the individual feature with the SVM classifier and the combined features with the SVM classifier. For this classifier, we used the linear polynomial kernel as the kernel function and the default values for all other parameters. We first performed 10-fold cross validation (CV) on the training set. For individual features, the best performance was obtained by both the semantic concept feature and the CEDD feature (with accuracy being around 97%). But the length of the semantic concept feature (which is 60) is much less than that of the CEDD feature (which is 144). The performance for classifying the images in the test set using the SVM model trained by using images in the training set was similar to that of the 10-fold CV. Table 3 presents the confusion matrix for the combined features for the 10-fold cross validation (CV) on the training set, and Table 4 presents the confusion matrix for the combined features for the testing set. In both tables, the corresponding precision and recall were both around 98%.

Table 2. Classification results (Features + SVM)

Feature	Dimension	Accuracy	
		10-fold CV	Testing set
Semantic concept	60	96.9%	96.8%
CEDD	144	96.4%	96.7%
FCTH	192	95.8%	96.1%
CLD	16	91.9%	92.5%
EHD	80	93.6%	92.4%
CEDD + FCTH + CLD + EHD	432	97.6%	97.8%
Semantic concept + CEDD + FCTH + CLD + EHD	492	98.1%	98.0%

Table 3. Confusion matrix (combined features + SVM) for 10-fold CV on the training set

Classified as →	Regular	Illustration
Regular	5894	106
Illustration	119	5881

Table 4. Confusion matrix (combined features + SVM) for testing set

Classified as →	Regular	Illustration
Regular	2163	37
Illustration	50	2150

We used the same training and test data to train and test CNN. Please note the network was not pre-trained with other data (for example, using ImageNet data). The number of epochs (one epoch = one pass through the training data) for CNN training was 150. Figure 5 shows the classification error rate on the training set and the test set as a function of the epoch number. The classification error rate for the test set was around 0.05 (i.e., classification accuracy was 95%) after 100 epochs. Table 5 lists the confusion matrix of the test set at epoch 150. The corresponding precision and recall were both around 95%.

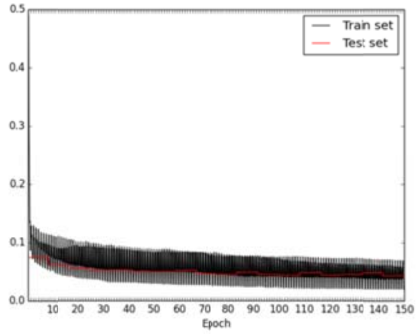


Fig. 5. CNN classification error rate

Table 5. Confusion matrix (CNN) for testing set

Classified as →	Regular	Illustration
Regular	2104	96
Illustration	93	2107

As demonstrated above, both methods achieved high performance for our figure modality classification application. The advantage of CNN is that it can avoid the process of manually identifying good features for a specific application. However, the conventional approach, i.e., handcrafting suitable and effective features, may achieve better performance, as demonstrated in these experiments. The experiments also demonstrate that the integration of semantic meanings into features is a promising way to reduce the semantic gap that frequently occurs when low-level visual features are used.

IV. CONCLUSION

In this paper, we presented our work on classifying figures in biomedical literature into two general classes: regular images and illustration images. This is the top level of the modality classification hierarchy utilized by Open-i, a multimodal search system that provides open access to nearly 3.2 million images from approximately 1.2 million Open Access biomedical research articles obtained from the NLM's PubMed Central (PMC) repository. We tested and compared two methods for this classification task. One is based on the conventional approach which includes feature extraction, by specifically identifying/applying (i.e., handcrafting) effective features. The other is based on deep learning, a relatively new technique for automatically learning representations of data from the raw pixel values. For the large dataset we tested which contains 16400 figures, both methods performed very well, achieving classification accuracy over 95%. The conventional method got a slightly better performance, which demonstrates the effectiveness of the features we chose.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

REFERENCES

[1] D. Demner-Fushman, S.K. Antani, M. Simpson M, G.R. Thoma, "Design and Development of a Multimodal Biomedical Information Retrieval System", JCSE, vol. 6, no.2, pp.168-177, June 2012.

[2] Alba García Seco de Herrera, Henning Müller and Stefano Bromuri, "Overview of the ImageCLEF 2015 medical classification task", in: CLEF working notes 2015, Toulouse, France, 2015

[3] E. Apostolova, D. You, Z. Xue, S.K. Antani, D. Demner-Fushman, G.R. Thoma, "Image Retrieval From Scientific Publications: Text and Image Content Processing To Separate Multipanel Figures", JASIST (Journal of the American Society for Information Science and Technology), 64(5):893–908, 2013.

[4] http://www.imageclef.org/2015/medical, access date: March 21, 2016

[5] C. Tsai, Bag-of-Words Representation in Image Annotation: A Review, ISRN Artificial Intelligence, Vol. 2012 (2012), Article ID 376804, 19 pages http://dx.doi.org/10.5402/2012/376804

[6] M. M. Rahman S. K. Antani, G. R. Thoma, A Medical Image Retrieval Framework in Correlation Enhanced Visual Concept Feature Space, 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS), August 3-4 2009, and Albuquerque, New Mexico, USA.

[7] V. Vapnik, Statistical Learning Theory. New York, NY, Wiley; 1998.

[8] T.F. Wu, C.J. Lin, R.C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling.", J. of Machine Learning Research, 5: 975–1005, 2004

[9] C. Chang and C. Lin, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27, 2011.

[10] S.A. Chatzichristofis, Y.S. Boutalis, “CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval,” In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) Proceedings of the 6th International Conference on Computer Vision Systems. Lecture Notes in Computer Science, vol. 5008, pp. 312-322, Springer- Verlag Berlin Heidelberg, 2008.

[11] S.A. Chatzichristofis, Y.S. Boutalis, “FCTH: Fuzzy color and texture histogram: A low level feature for accurate image retrieval,” In: Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 191-196, 2008.

[12] M. Lux, “Caliph & Emir: MPEG-7 photo annotation and retrieval,” Proceedings of the seventeen ACM international conference on Multimedia, pp. 925-926, 2009, Beijing, China.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, “The WEKA Data Mining Software: An Update”, SIGKDD Explorations, 11 (1), 2009.

[14] LeCun, Y., Bengio, Y., Hinton, G., “Deep learning”, Nature, 521, pp. 436-444, May 2015.

[15] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 86 (11): 2278–2324, November 1998.

[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., “ImageNet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision, 2015.

[17] https://code.google.com/p/cuda-convnet/, access date: March 21, 2016

[18] Krizhevsky, A., Sutskever, I., Hinton, G., “ImageNet classification with deep convolutional neural networks,” Neural Information Processing Systems (NIPS), pp.1097-1105, 2012.

[19] http://www.ncbi.nlm.nih.gov/pubmed, access date: March 21, 2016