Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference

Exploring Interaction Between Images and Texts for Web Image Categorization

Lei Li[1*], **Wenting Lu**[2†], **Jingxuan Li**[1], **Tao Li**[1] **Honggang Zhang**[2], **Jun Guo**[2]
[1]School of Computing and Information Sciences
Florida International University
Miami, Florida 33199, United States
[2]Pattern Recognition and Intelligence System Lab
Beijing University of Posts and Telecommunications
Beijing 100876, P.R.China

Abstract

With the rapid development of technologies for fast access to the Internet and the popularization of digital cameras, enormous digital images are posted and shared online everyday. Simultaneously, web images are usually organized by topics of events and are often assigned appropriate topic-related text descriptions. Given a set of images along with corresponding texts, a challenging problem is how to utilize the available information to perform image retrieval tasks, such as image classification and image clustering. Previous works on image categorization focus on either adopting text or image features, or simply combining these two types of information together. In this paper, we propose two novel approaches (**Dynamic Weighting** and **Region-based Semantic Concept Integration**) to categorize the images under the "supervision" of topic-related text descriptions; In addition, we provide a comparative experimental investigation on utilizing text and image information to tackle image classification. Empirical experiments on a manually collected image dataset (consisting of images related to the events after disasters) demonstrate the efficacy of our proposed classification methods.

Introduction

Multimedia information plays an increasingly important role in human's daily activities. With the rapid development of technologies of fast access to the Internet and the popularization of digital cameras, enormous digital images are posted and shared online everyday. Besides great convenience, how to retrieve images that satisfy the needs of web users in multimedia databases is becoming more and more difficult and challenging. Particulary, web image categorization, as a crucial step of image retrieval, attracts much more attention and is very useful in the subsequent procedures, such as indexing and organizing web image databases, browsing and searching web images, and discovering interesting patterns from images (Yin et al. 2009).

However, web image categorization is not a trivial task due to the diversity of image content and the limited information. In general, the images posted on the Internet may have great visual differences, rendering image categorization challenging since it is difficult to extract common features shared by most images as the comparison base. Fortunately, text information is often provided by web users to describe the general contents of images, *e.g.*, image titles, headers, or text descriptions assigned to them. A possible solution to web image categorization is to take such text information into consideration. Specifically, we can initially extract different types of features (image and text features), and then categorize the images by delving into the special characteristics of the integrated version of different features.

In this paper, we explore the feasibility of using text information as a "guidance" for image categorization by proposing two novel methods (**Dynamic Weighting** and **Region-based Semantic Concept Integration**), which achieve better performance comparing with existing approaches. Specifically, the proposed *Dynamic Weighting* assumes that different image features might carry semantic meanings with different significance, and under the "supervision" of text features, the importance of different image features can be dynamically decided. It is straightforward that the important features may have more distinguishable power for image categorization. Another proposed method *Region-based Semantic Concept Integration* first segments images into different regions, and then categorizes images based on the correlation between regions and semantic concepts. Moreover, we provide a comparative experimental study on integrating text and image information to perform image categorization. Empirical experiments on a manually collected image dataset (including images related to the events after disasters) demonstrate the efficacy of our proposed methods.

The rest of this paper is organized as follows. In Section 2 we review some related works that combine image and text features to perform image classification tasks. In Section 3 we give algorithmic details of the two proposed approaches to effectively integrate text information with image information for image categorization. Section 4 presents a detailed experimental comparison among different approaches and finally we conclude the paper in Section 5.

Related Work

Most of the existing web image categorization approaches often focus on utilizing text descriptions of images to categorize images via simply matching keywords. Currently, the majority of web search engines still adopt this technique due

[*]Li, L. and Lu, W. contributed equally to this work.
[†]This author is a visiting Ph.D student.

to its fast speed. There are some limitations in text-based image categorization: (1) web images cannot be appropriately classified if there is no text information assigned to them; (2) the manual text labeling is too subjective due to human assignments, which might result in bias or noise to web image categorization; (3) using a few words to describe the content of an image is not enough since the limited text description can only provide a relatively sparse feature space. Therefore, the performance of traditional text-based web image categorization systems is very limited.

To solve the above problems, many research publications (Blei and Jordan 2003; Giacinto, Roli, and Fumerga 2002; Kalva, Enembreck, and Koerich 2007; Wu et al. 2004; Zhu, Yeh, and Cheng 2006) aim to design multi-view learning algorithms to learn classifiers from multiple information sources via integrating different types of features together to perform classification. In general, multi-view learning methods can be categorized into three different groups:

1. Feature Integration: Enlarge the feature representation to incorporate all attributes from different sources and produce a unified feature space. The advantage of feature integration is that the unified feature representation is often more informative and also allows many different data mining methods to be applied and systematically compared. One disadvantage is the increased learning complexity and difficulty as the data dimension becomes large (Wu, Oviatt, and Cohen 2002);
2. Semantic Integration: Keep data intact in their original form and computational methods are applied to each feature space separately. Results on different feature spaces are then combined by either voting (Carter, Dubchak, and Holbrook 2001), Bayesian averaging (Bishop 2006), or the hierarchical expert system approach (Jordan and Jacobs 1994). One advantage of semantic integration is that it can implicitly learn the correlation structure between different sets of features (Li and Ogihara 2005).
3. Intermediate Integration: A compromise between the feature integration and the semantic integration. The idea is to keep the feature spaces in their original forms and integrate them at the similarity computation or the Kernel level (Schölkopf and Smola 2002; Lanckriet et al. 2004). Different weights can be assigned different data sources. Standard computational methods can then be applied once the total similarity is computed.

Our contribution: In order to explore the feasibility of using texts as a guidance for image classification, we propose two novel multi-view learning methods to achieve better performance for web image categorization. Further, we present an empirical investigation on different methods for combining text information with image features and compare their classification performance.

Classification Algorithms

Web image categorization is a key step for many web-based multimedia applications. It is crucial for the subsequent processes (e.g., image retrieval) and has a direct impact on the speed and accuracy of other applications related to images on the web. As mentioned above, three different multi-view learning approaches have been used to resolve the problem of web image categorization. However, all these approaches focus on simply combining two data sources (text and image information), and none of them takes the advantage that one data source can provide "guidance" for another on how to perform categorization task. In this paper, besides providing a comparative study of the previous works, we also propose two novel web image categorization methods – *Dynamic Weighting* and *Region-based Semantic Concept Integration* – which employ the text-based information (*i.e.*, the text itself and the semantic concepts hidden in the text) to guide the classification, and consequently achieve better image categorization results.

Dynamic Weighting

Image feature extraction techniques tend to extract a huge number of image features based on different criteria. Among these features, some of them might carry significant semantic information about the image, whereas some others might be less crucial for the tasks being executed on the image. Particularly in image classification, the extracted features should be more representative and carry substantial amount of semantic meanings. Therefore, it might be helpful to dynamically assign different weights to different image features so that the features with more importance can be captured and play more meaningful roles on the classification. Some previous works (Shao et al. 2009) on music information retrieval demonstrate how to learn appropriate similarity metrics based on the correlation between acoustic features and user access patterns. Motivated by this, we incorporate the concept of dynamic feature weighting into our image classification problem.

Specifically in image classification, given that human perception of an image is well approximated by its text description, a good weighting schema for the extracted image features guided by text information may lead to a good similarity measurement, and therefore better classification results. Let $\mathbf{m}_i = (\mathbf{f}_i, \mathbf{t}_i)$ denote the *i*-th image in the image collection, where $\mathbf{f}_i$ and $\mathbf{t}_i$ represent its image features and text features respectively. Let $S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w}) = \sum_l f_{i,l} f_{j,l} w_l$ be the image-based similarity measurement between the i-th and the j-th images when the parameterized weights are given by $\mathbf{w}$. Let $S_t(\mathbf{t}_i, \mathbf{t}_j) = \sum_k t_{i,k} t_{j,k}$ be the similarity measurement between the i-th and the j-th images based on their text description features, in general, the words with specific meanings extracted from texts. Here for each k, $t_{i,k}$ denotes whether the k-th word appears in the text description of the i-th image. To learn appropriate weights $\mathbf{w}$ for image features, we can enforce the consistency between similarity measurements $S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w})$ and $S_t(\mathbf{t}_i, \mathbf{t}_j)$. The above idea leads to the following optimization problem:

$$\mathbf{w}^* = argmin \sum_{i \neq j} \left(S_f(\mathbf{f}_i, \mathbf{f}_j; \mathbf{w}) - S_t(\mathbf{t}_i, \mathbf{t}_j)\right)^2 \quad s.t. \mathbf{w} \geq 0. \tag{1}$$

Let p be the number of image features. The summation in

Eq.(1) can be rewritten as follows:

$$\sum_{i\neq j}(S_f(\mathbf{f}_i,\mathbf{f}_j;\mathbf{w})-S_t(\mathbf{t}_i,\mathbf{t}_j))^2$$
$$= \sum_{i\neq j}\left(f_{i,1}f_{j,1}w_1+\cdots+f_{i,p}f_{j,p}w_p-\sum_k t_{i,k}t_{j,k}\right)^2$$
$$= \sum_{i\neq j}\Bigg((f_{i,1}f_{j,1}w_1+\cdots+f_{i,p}f_{j,p}w_p)^2 -2(f_{i,1}f_{j,1}w_1+\cdots+f_{i,p}f_{j,p}w_p) \times\left(\sum_k t_{i,k}t_{j,k}\right)+\left(\sum_k t_{i,k}t_{j,k}\right)^2\Bigg),$$

where $f_{i,l}$ is the l-th feature in the image feature set f_i and $f_{j,l}$ is the l-th feature in the image feature set f_j. Let n be the number of images, and let

$$F=\begin{bmatrix} f_{1,1}f_{2,1} & f_{1,2}f_{2,2} & \cdots & f_{1,g}f_{2,g}\\ \cdots & \cdots & \cdots & \cdots\\ f_{n-1,1}f_{n,1} & f_{n-1,2}f_{n,2} & \cdots & f_{n-1,g}f_{n,g}\end{bmatrix},$$

and

$$T=\begin{bmatrix}\sum_{i\neq j} f_{i,1}f_{j,1}(\sum_k t_{i,k}t_{j,k})\\ \vdots\\ \sum_{i\neq j} f_{i,g}f_{j,g}(\sum_k t_{i,k}t_{j,k})\end{bmatrix},$$

where F is a $(\binom{n}{2}\times p)$ matrix and T is a $(p\times 1)$ matrix. Thus, Eq.(1) is equivalent to

$$\mathbf{w}^* = argmin\left[\frac{1}{2}\times 2(Fw)^T(Fw)-T^Tw\right]$$
$$= argmin\left[\frac{1}{2}(w^T(2F^TF)w+(-2T^T)w)\right] s.t. \mathbf{w}\geq 0.$$

This optimization problem can be addressed using quadratic programming techniques (Gill, Murray, and Wright 1981). After calculating the dynamic weights for each image features, we multiply the feature values with the corresponding weights for each image, and finally obtain a new feature space with weighted information. These feature vectors can then be fed into classifiers. In the experiments, we will show how much the classification results are influenced by our dynamic weighting schema.

Region-based Semantic Concept Integration

In the real-world applications, an image always contains various semantic concepts and these concepts often intersect with each other, which is not helpful to efficiently extract semantic information. In this section, we explore the feasibility of utilizing the underlying semantic concepts of text information as a "guidance" to facilitate image categorization. To address the issue mentioned above, we firstly divide original images into different regions to ensure that the content of each region represents almost the same local pattern, and then based on the local semantic patterns of the images, we propose our *Region-based Semantic Concept Integration* method. Figure 1 shows the framework of our proposed approach, which can be divided into four different

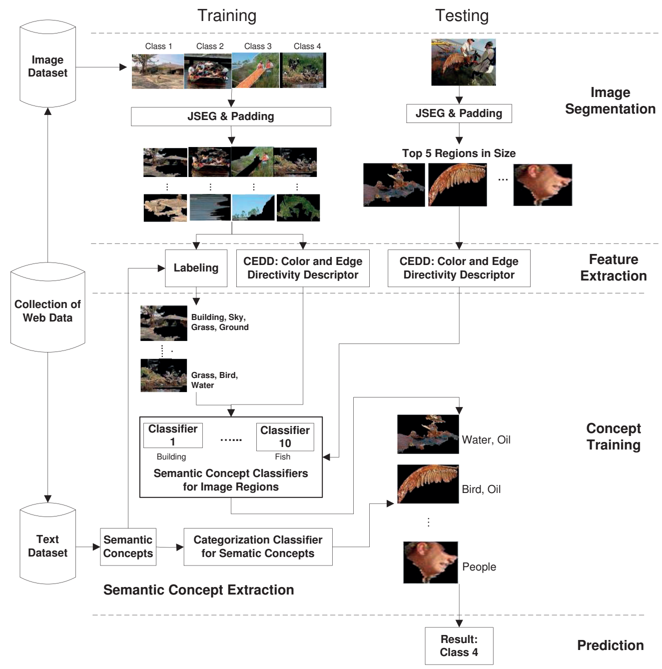

Figure 1: Framework of region-based concept integration.

sub-processes: *semantic concept extraction*, *image segmentation*, *feature extraction* on each region, and *region-based semantic concept classification*. In the following, we provide algorithmic details of these four processes.

Semantic Concept Extraction In this process, we initially analyze the text description related to each image, and then obtain some original high-frequency terms in these texts by using MALLET (McCallum 2002), a java-based package for statistical natural language processing. We then compare the semantics of these high-frequency words and summarize them to generate several most general semantic concepts using WordNet (Miller 1995). The general concepts are represented as the hypernyms of the high-frequency words. Then each text description can be represented by the combination of these concepts. The generalized concepts can provide guidance on how to select image region samples in the training step of semantic concept classifier, as well as to train a concept model that builds the relationship between semantic concepts and original categories, which will be described in Section "Design of Semantic Concept Classifiers".

Image Segmentation In order to associate the images with the generalized concepts extracted from the procedure of "Semantic Concept Extraction", we need to segment images into different regions such that each region can be related to one or more semantic concepts. Ideally, image segmentation aims to divide original web images into different regions based on the criterion that each region contains only one object or one part of an object. However, due to the limitation of current segmentation techniques, it is very difficult to perfectly segment images (Liu, Zhang, and Lu 2008;

Li, Socher, and Fei-Fei 2009). In this paper, we use a state-of-the-art segmentation method JSEG(Joint Systems Engineering Group) (Deng and others 2001), which segments images based on color and texture information. In the segmentation algorithm, image color space is first quantized into several classes, and a color class-map of each image is then obtained via re-representing each pixel of the original image by its corresponding color class label. After that, the spatial segmentation is performed on this color class-map which can be viewed as a special type of texture composition. Here, a criterion named "*J*-Value" is used to measure whether the segmentation is reasonable or not. If one image consists of several homogeneous color regions, the color classes will be separated from each other and the value of *J* is larger. Figure 2 presents some examples of the segmented regions of web images using JSEG.

Figure 2: An illustration of image segmentation.

Region Feature Extraction After segmenting images into different regions, image feature extraction is performed on each region. Note that color and texture are two of the most general global features in the field of image processing and computer vision, both of which have their own advantages and drawbacks. In order to represent image/region effectively, we adopt CEDD (Color and Edge Directivity Descriptor) (Chatzichristofis and Boutalis 2008), which is a new low-level feature descriptor incorporating both color feature and texture feature. In CEDD, a novel but effective method is adopted to integrate a 24-bins color histogram and a 6-bins texture histogram to form a final 144-bins histogram. One of the most important characteristics of CEDD is its low computational power needed for feature extraction, in comparison with the needs of most of MPEG-7 descriptors. For detailed algorithmic procedure of CEDD, please refer to (Chatzichristofis and Boutalis 2008).

Design of Semantic Concept Classifiers In this step, the design of semantic concept classifiers can be divided into the following two parts:

Semantic Concept Classifiers for Image Regions: Based on previous steps of image segmentation and feature extraction, we obtain a set of regions and their corresponding 144-dimension feature vectors. Then we need to train *N* (the number of generalized semantic concepts obtained from text descriptions) different semantic concept classifiers respectively using training regions. Each of the training regions is manually labeled with multiple concepts, but each semantic concept classifier is just designed for binary classification identifying whether a region contains this concept or not. Therefore, when a testing region arrives, it will be fed into these *N* classifiers respectively and a *N*-dimension vector is obtained for the input of the categorization classifier.

Categorization Classifier for Semantic Concept: We design a multi-label classifier to build the relationship between semantic concepts and the original categories. Here we use One-Against-One (OAO) (Hsu and Lin 2002) method to design the multi-label SVM classifier. OAO designs an original binary SVM classifier between two random classes of samples, and it needs $k(k-1)/2$ original binary SVM classifiers. In addition, we use a voting strategy in classification, in which each binary classification is considered to be a voter where votes can be cast for all the regions, and finally each region is designated to be in a class with the maximum number of votes.

When a testing image comes, it will be segmented into several regions. Then the system will choose the top *M* regions in size automatically (if the actual number of regions for the image is less than *M*, the system will adopt the actual number of regions) and extract CEDD feature from these regions. Note that the larger the region is, the more information it contains, and if the region is too small, it is difficult to identify the exact semantic meaning and therefore may involve noises. After that, each region will pass through the *N* different semantic concept classifiers respectively, to identify whether this region contains the concept or not. If the region contains this concept, a concept label will be assigned to this region so that each region will be assigned multiple concept labels, and the region can be represented by a *N*-dimension vector. Then we integrate these *M* *N*-dimension vectors into a single *N*-dimension vector describing whether the original image contains certain concepts or not. Finally, we use the categorization classifier for semantic concepts to predict which class the original testing image belongs to.

Experiments

Real World DataSet

Unlike the traditional scene object databases, which are mainly focusing on visual categorization, web images are usually organized by topics of events. From a classification perspective, the differences between these two kinds of images are as follows:

- Images in the same category for visual scene categorization are visually similar, but object scaling, rotating, occluding and submerging often happen in clutter background; Comparatively, web images in the same category may vary visually but very similar in terms of semantic concepts;
- Images in the same visual category contain the same object or scene and it would occupy most area of the whole image, whereas web image focus on reflecting just one aspect of the whole event;

Therefore in our work, we focus on the conceptual information contained in images, but not simple rotation or zooming on the same objects.

Dataset Discription To systematically study the previous multi-modal feature combination methods and compare them with our proposed approaches, we manually collected 355 colored web images and their corresponding text descriptions about "the aftermath of disasters", which include 4 different topics: Hurricane building collapse, Hurricane flood, Oil spill seagrass, and Oil spill animal death. Each category includes 101, 101, 53 and 100 images respectively, and the entire image dataset is split into two parts: about 70% images (247 in total) are randomly selected for training and the rest 30% images (108 in total) are taken as the test data.

Note that for *Region-based Semantic Concept Integration* method, we generalize 10 (*i.e.*, *N*=10) concepts from the text description set, including "building", "water", "sky", "grass", "oil", "bird", "ground", "people", "helicopter" and "fish". In addition, we generate 1573 regions from 247 training images with the guidance of these semantic concepts, and 513 regions from 108 testing images by automatically choosing the top 5 (*i.e.*, *M*=5) regions in size (if the actual number of regions is less than 5, the system will adopt the actual number of regions).

Design of Experiments

In our experiment, we use LIBSVM (Chang and Lin 2001) as our base classification tool. The parameter tuning is done via *k*-fold cross validation. For the purpose of comparison, we first implement five existing methods for web image categorization, then compare their classification performance with those of our proposed approaches. These five existing methods include:

- Text-based Classification (*Text* for short): Extract text features from texts assigned to the corresponding images, and then use these features to feed SVM classifier;
- Image-based Classification (*Img* for short): Extract CEDD image features from the images, and then use these features to feed SVM classifier;
- Feature Integration (*Feat* for short): Treat unique terms as text features and extract image features using CEDD. Note that the extracted CEDD feature is a 144-dimension vector, while the cardinality of the text features is 1788. To balance the contribution of different features to the classification results, we choose the top 144 terms with high frequency as the text features. We combine the features of text and image together by simply concatenating these two types of features to form a 288-dimension vector as the input of SVM classifier.
- Semantic Integration (*Sem* for short): Train two classifiers based on text features and image features (CEDD) respectively, and then ensemble these two classifiers to be an integrated version, similar to the method proposed in (Carter, Dubchak, and Holbrook 2001).
- Intermediate Integration (*Sim* for short): Compute the pairwise similarity using text-based features and image-based features (CEDD) respectively, and then use the weighted summation of these two types of similarities as the similarity measurement between images. Note that different weights can be assigned to the features of different data sources. We tune the weight factor to find the optimal one through empirical comparison.

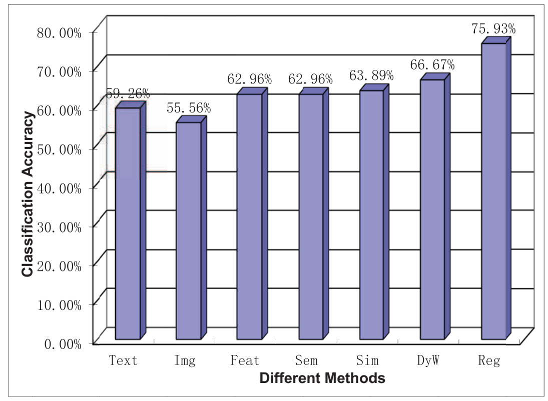

Figure 3: Comparison among the five existing methods and our two proposed methods based on the accuracy of web image categorization.

Experiment Results

Comparison among all the methods In Figure 3, the comparison among the above 5 different methods and our proposed methods (*DyW* and *Reg* for short) based on the accuracy of web image categorization results are presented.

From the comparison results, we observe that the best performance of web image categorization using single-modal approaches is less than 60%. However, once the text and image data sources are integrated using two-modal information fusion techniques, the categorization performance is improved. The intuitive explanation for the improvement is that two-modal approaches are able to incorporate the advantages of the two data sources together, which leads to better categorization results compared with using only one type of features (here the "advantage" represents the positive contribution of the features of certain data source to the image categorization results). In addition, compared with the categorization results provided by the five existing methods, our proposed approaches outperform others. The reason behind the performance improvement is straightforward: our two proposed methods share one common characteristic – employing advantages of one data source to enrich the other data source. In other words, these two methods explore inherent connections between two data sources by utilizing text-based information to either find out the best weighting schema for the image-based features or generate the classifiers from the semantic image regions related to text concepts.

Comparison between *DyW* and *Reg* We further compare two proposed approaches based on their classification performance on each category of images. The comparison results are shown in Table 1. From the results, we have the following observations:

	Dynamic Weighting			Region-based Concept		
	Precision	Recall	F1-measure	Precision	Recall	F1-measure
Category 1	0.4565	0.6774	0.5455	0.6429	0.8710	0.7397
Category 2	0.6875	0.7097	0.6984	0.7391	0.5484	0.6296
Category 3	1.0000	0.0625	0.1176	1.0000	0.5000	0.6667
Category 4	0.9655	0.9333	0.9492	0.8571	1.0000	0.9231

Table 1: Comparison of classification results on each category using our proposed methods. Note that "Category 1-4" represents Hurricane building collapse, Hurricane flood, Oil spill seagrass, and Oil spill animal death respectively.

1. *DyW* and *Reg* provide reasonable performance on Category 1 and 2, and good results on Category 4.
2. The performance of these two methods on Category 1 and 2 is slightly worse than the results on Category 4. This is due to the characteristics of our image dataset. Most of the images in Category 1 and 2 contain a lot of " semantic noise". For instance, most of the images in Category 1 focus on the collapse "buildings", but "grass" and "water" also appear in these images. These "noise" would cause our classification methods to misclassify these images into the other categories. Even though *Reg* incorporates the semantic information and shows better results than *DyW*, "noise" still exists to some extent.
3. The recall of both two methods on Category 3 is very low. After analysis, we found that most of the images in Category 3 are about "grass"; however, "grass" appears in almost all the categories, which results in the misclassification of the images.
4. *Reg* outperforms *DyW* on category 1 and 3, and the performance of *Reg* on category 2 and 4 is comparable with *DyW*. The reason that the overall performance of *Reg* is better than the one of *DyW* is that *Reg* could benefit from the semantic information hidden in the text whereas *DyW* only make use of raw text information.

Conclusions

In this paper, we study the problem of combining two data sources (text and image) to perform image categorization tasks and show that such combination can lead to better classification results comparing with using individual data sources. Also, we propose two novel multi-view learning methods which can effectively utilize the image-related text data to find out better schemas to classify the images. The empirical results show that our proposed methods outperform the previous methods in terms of the accuracy of classification results, and they can provide solid basis for the subsequent procedures of image retrieval.

Acknowledgement

This work is partially supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and by the Army Research Office under grant number W911NF-10-1-0366, and by the Fundamental Research Funds for the Central Universities, 111 Project of China (B08004) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. We would also like to thank the Distributed Multimedia Information Systems Laboratory at Florida International University for helping us collect the dataset.

References

Bishop, C. 2006. *Pattern recognition and machine learning*.

Blei, D., and Jordan, M. 2003. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 127–134. ACM.

Carter, R.; Dubchak, I.; and Holbrook, S. 2001. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Research* 29(19):3928.

Chang, C., and Lin, C. 2001. LIBSVM: a library for support vector machines.

Chatzichristofis, S., and Boutalis, Y. 2008. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Proceedings of the 6th International Conference on Computer Vision Systems*, 312–322.

Deng, Y., et al. 2001. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 800–810.

Giacinto, G.; Roli, F.; and Fumerga, G. 2002. Unsupervised learning of neural network ensembles for image classification. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, volume 3, 155–159. IEEE.

Gill, P.; Murray, W.; and Wright, M. 1981. Practical optimization.

Hsu, C., and Lin, C. 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13(2):415–425.

Jordan, M., and Jacobs, R. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation* 6(2):181–214.

Kalva, P.; Enembreck, F.; and Koerich, A. 2007. Web image classification based on the fusion of image and text classifiers. In *Proceedings of the 9th International Conference on Document Analysis and Recognition*, 561–568. IEEE Computer Society.

Lanckriet, G.; Cristianini, N.; Bartlett, P.; Ghaoui, L.; and Jordan, M. 2004. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research* 5:27–72.

Li, T., and Ogihara, M. 2005. Semisupervised learning from different information sources. *Knowledge and Information Systems* 7(3):289–309.

Li, L.; Socher, R.; and Fei-Fei, L. 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2036–2043. IEEE.

Liu, Y.; Zhang, D.; and Lu, G. 2008. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition* 41(8):2554–2570.

McCallum, A. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Miller, G. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11):39–41.

Schölkopf, B., and Smola, A. 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Shao, B.; Ogihara, M.; Wang, D.; and Li, T. 2009. Music recommendation based on acoustic features and user access patterns. *IEEE Transactions on Audio, Speech, and Language Processing* 17(8):1602–1611.

Wu, Y.; Chang, E.; Chang, K.; and Smith, J. 2004. Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, 572–579. ACM.

Wu, L.; Oviatt, S.; and Cohen, P. 2002. Multimodal integration-a statistical view. *IEEE Transactions on Multimedia* 1(4):334–341.

Yin, Z.; Li, R.; Mei, Q.; and Han, J. 2009. Exploring social tagging graph for web object classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 957–966. ACM.

Zhu, Q.; Yeh, M.; and Cheng, K. 2006. Multimodal fusion using learned text concepts for image categorization. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, 211–220. ACM.