Meta-classifiers For Multimodal Document Classification

Scott Deeann Chen [#1], Vishal Monga [*2], and Pierre Moulin [#3]

[#] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign*
405 North Mathews Avenue, Urbana, IL 61801 USA
[1] chen124@illinois.edu
[3] moulin@ifp.uiuc.edu

[*] *Xerox Research Center Webster*
800 Phillips Road, Webster, NY, 14580, USA
[2] Vishal.Monga@xerox.com

***Abstract*—This paper proposes learning algorithms for the problem of multimodal document classification. Specifically, we develop classifiers that automatically assign documents to categories by exploiting features from both text as well as image content. In particular, we use meta-classifiers that combine state-of-the-art text and image based classifiers into making joint decisions. The two meta classifiers we choose are based on support vector machines and Adaboost. Experiments on real-world databases from Wikipedia demonstrate the benefits of a joint exploitation of these modalities.**

I. Introduction

Document classification, or the task of automatically assigning semantic categories to electronic documents, has been actively researched over the last two decades. Apart from traditional applications in document database management, many new applications have emerged in the context of email spam filtering, and web-page classification.

As with any classification problem, document classification is comprised of two stages: 1.) a feature extractor, and 2.) a decision stage that actually performs the assignment of documents to classes based on the extracted features. Several text feature extractors have been proposed [1], but by far the most popular ones have been variants of the term-frequency vector by Salton *et al.* [2]. A wealth of data mining and machine learning techniques have then been applied to and/or developed for the purposes of document classification. These include the naive Bayes classifier, k-nearest neighbor classifier, neural networks, decision trees, logistic regression and most recently support vector machines (SVM) [3]. While the aforementioned advances have been significant both conceptually and from the viewpoint of enhancing classification accuracy, it is evident that document classification methods have largely focused on intelligent mining of text data in the documents.

An increasingly large number of document collections however contain significant imagery. Examples include official documents such as annual reports, advertisement brochures, technical and scientific articles, and organized web-page collections such as Wikipedia. Often the image data convey information about the document category and show high correlation with the text content. For example, an advertisement brochure around digital print products will speak to offerings in print-capabilities such as printer design, color image quality, pages-per minute and include corresponding images of the actual printer, sample prints, color charts etc. Further, psychological studies [4] on the contribution of multimodal data in documents such as animation, images, text has confirmed better comprehension by human readers as opposed to a single modality.

Motivated by this observation, this article explores the use of image-based features and in particular their combination with the well-established body of algorithms in using text features for the purposes of *multimodal* document classification. Multimodal classification is of course a very active area of research for a variety of other problems, e.g. the joint use of audio and visual clues for video classification [5]. Some of these approaches have also recommended using text from meta-data tags and captions as one of the modalities [6] to enhance classification performance. Recently, using figure captions and meta-data text in conjunction with image features has shown promise in enhancing image search and classification [7], as well as clustering of web pages [8].

This paper augments these recent observations by taking a multimodal signal processing approach to document classification. First, we employ state-of-the-art text and image feature extraction and design classifiers that can assign a multimedia document to pre-determined categories based on features individually from each of the two modalities. Next, meta-classifiers are proposed that combine classifier decisions from the individual modalities. In all generality, a meta-classifier aims to combine the merits of multiple learning algorithms to improve the overall performance. While truly optimal meta-classification is in general an open problem, we develop two intuitively motivated meta-classification algorithms: 1.) a SVM based meta-classification [9] where soft measures from stage 1 of text and image classification are used as features in stage 2, and 2.) AdaBoost [10] which iteratively boosts several *weak* text and image feature based learners into a strong one.

To validate our claims of the benefits of integrating image data, we test on *real-world* multimedia document pages acquired from Wikipedia. Experiments on four distinct document

collections reveal that even with rather pedestrian classification results using a single modality, significant improvements can be obtained in classification accuracy by joint use of text and image features. Further, the benefits of joint use of the modalities are observed for both SVM as well as AdaBoost based meta-classification where one may be preferred to the other based on performance for particular classes or varying precision-recall rates.

II. Background

A. Text Feature Extraction

The most popular text document feature extraction method is bag-of-words [3], which treats each single word as a feature. The value of each feature for a document is the number of occurrences of the corresponding word. Also, a dictionary is defined as the collection of words contained in the document. For a set of d documents and a dictionary with t terms, a term-frequency matrix is defined as a $d \times t$ matrix $\mathbf{A} \in R^{d \times t}$. Each element $A(i,j)$ in A, where $i \in \{1,2,\ldots,d\}$ and $j \in \{1,2,\ldots,t\}$, is defined as the measurement of the frequency of the j^{th} word of the dictionary in the i^{th} document. The text feature vector for the i^{th} document is the i^{th} row of A.

However, not all words contain semantic information about the document. To remove them, two main techniques are used. One method is stop-words elimination [1], which observes that words such as "the" or "and" convey no meaningful information about the document. Another method is document-frequency thresholding, which rejects infrequent words because they convey little information. In our work, we apply these two techniques to the term-frequency matrix to reduce the dimension of text feature vectors.

B. Image Feature Extraction

Bag-of-features is a technique which uses local features to represent an image. It has been found to be more useful than global features in image classification [11]. The underlying idea is similar to bag-of-words, called "visual words".

There are however three key differences between *visual* words and *regular* words. First, unlike bag-of-words, which uses every word in the document, not every pixel forms a visual word. Only visually significant points of interest are taken into account. Such points of interest have been actively researched in computer vision and are typically determined using geometrically meaningful feature point detectors. Remarkably, random sampling techniques have also shown to be useful for determining feature points [12]. Such techniques essentially select significant neighborhoods or patches in the vicinity of the feature point.

Second, for performance reasons, the sampled patches are not directly used but represented by descriptors [13], [14]. Scale-invariant feature transform (SIFT) descriptors are known to be amongst the most competitive [14] and were employed in our work along with random sampling strategies.

Finally, there is no pre-determined dictionary for visual words. Usually, a dictionary is created via a training process using unsupervised learning algorithms [12]. Each patch is then assigned the visual word that corresponds to its nearest neighbor in the dictionary.

The feature vector generation process is then straightforward. For a set of d images and a dictionary with t terms, a feature-frequency matrix is defined as a $d \times t$ matrix $\mathbf{B} \in R^{d \times t}$. Each element $B(i,j)$ in B, where $i \in \{1,2,\ldots,d\}$ and $j \in \{1,2,\ldots,t\}$, is defined as the measurement of the frequency of the j^{th} visual word of the dictionary in the i^{th} image. The image feature vector for the i^{th} image is the i^{th} row of B.

III. Proposed Classification Algorithms

As described in Section II, images differ greatly from texts both in nature and dimensionality. That is, both image and text features carry information about the document category but their correlations are not readily apparent in their native (low-level) form.

In view of the above, we employ a two-stage meta-classification approach [6]. The first stage synthesizes a *meta-feature vector* which is obtained by using "soft-outputs" from individual text and image classifiers. The second stage makes the assignment decision to a document category based on the meta-feature vector. There are two merits to such an approach: 1.) Stage 1 ensures that information from these distinct modalities is brought to a *common ground*, and 2.) Because stage 2 observes outputs of several image and text classifiers, it has more information at hand to make better decisions. The two meta-classifiers we propose are described next.

A. Meta-classification with Support Vector Machines

Support Vector Machines (SVM) based meta-classification is illustrated in Fig. 1.

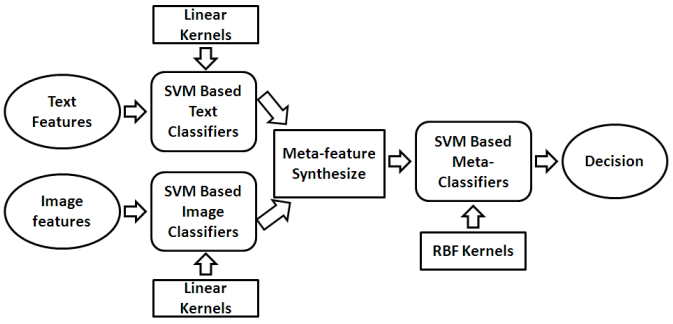

Fig. 1. SVM Based Meta-Classification

SVM is a widely used learning technique which finds the margin-maximizing hyperplane in the feature space [15]. Margin maximization has strong theoretic backing [16] in structural risk minimization which aims to bound the generalization error of the classifier. The general form of the decision function f of a SVM classifier may be written as:

$$f(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i . y_i . K(\mathbf{s}_i, \mathbf{x}) + b \qquad (1)$$

where $\mathbf{x}$ is the multi-dimensional feature vector, $\mathbf{s}_i$ are the *support* vectors, N_s the number of support vectors, and $y_i \in \{1, -1\}$ the corresponding classifier decisions. The positive Lagrange multipliers α_i's and hyperplane parameter b are determined by solving the constrained optimization problem that manifests itself in margin maximization. A binary decision of $\{1, -1\}$ is made depending on whether $f(\mathbf{x}) > 0$ or not. Soft-output decisions that depend on the exact numerical value of $f(\mathbf{x})$ are also popular and can convey the "confidence" of the classifier.

Given a set of training documents, the text and image features are first extracted using the techniques mentioned in II-A and II-B respectively. For a document database with m categories, m base SVM classifiers for text and m base SVM classifiers for images are trained in an one-against-all manner [17]. This means that a vector of dimension m which comprises of soft-outputs is obtained from each set of classifier.

For a given document, let $\mathbf{x}_t \in \mathbf{R}^m$ denote the vector of soft-outputs from the text-based SVM. An interesting challenge presents itself in constructing the corresponding vector from images. This is because the number of images in each document may vary. We therefore need a strategy to get a single vector $\mathbf{x}_i$ that is representative of *all* images in the document.

Given a document with p images, one may obtain soft-output vectors $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}$ from p SVMs applied to the features from p images. We propose two strategies to compute $\mathbf{x}_i$. These strategies are summarized in Table I. The first one simply computes $\mathbf{x}_i$ as an average vector of soft-outputs $\mathbf{x}_i = \sum_{l=0}^{p} \mathbf{x}_{il}$. The implicit assumption is that each image in the document has the same importance.

The above may of course not always be true. Instead, as is observed in practice, a *dominant* image type may appear in documents that belong to the same category. This situation is handled using our second strategy in Table I. We first apply k-means clustering [18] on $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}$. Let k be the number of clusters, $\mathbf{x}_{\mu 1}, \ldots, \mathbf{x}_{\mu k}$ be the mean of the clusters and N_j, $j = 1, 2, .., k$, be the number of images that belongs to the k^{th} cluster. The second strategy takes $\mathbf{x}_i = \mathbf{x}_{\mu k'}$ as the representative feature vector, where $k' = argmax_k N_k$.

The meta-feature vector for the document $\mathbf{x}_m$ is then defined as the concatenation of $\mathbf{x}_t$ and $\mathbf{x}_i$ and the stage 2 meta-classifiers are trained on these meta-features.

In employing SVMs, often the choice of kernel $K(\mathbf{s}_i, \mathbf{x})$ is important and application/feature dependent. In this work, we employ linear kernels on individual text and image features in Stage 1 (see Fig. 1) because they are high-dimensional and sparse [3]. Radial based function (RBF) kernels were used on the low-dimensional meta-feature vector in stage 2.

TABLE I
STRATEGIES FOR COMBINING META-FEATURES FROM MULTIPLE IMAGES

Strategy 1: Average
Given: A set of vectors $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}$.
Output: $\mathbf{x}_i = \sum_{l=0}^{p} \mathbf{x}_{il}$

Strategy 2: Mean of Largest Cluster
Given: A set of vectors $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}$.
Step1: Apply k-means clustering to $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{ip}$.
Step2: Find the mean of the largest cluster $\mathbf{x}_{\mu k'}$.
Output: $\mathbf{x}_i = \mathbf{x}_{\mu k'}$

B. Meta-Classification Via Adaboost

Boosting has its roots in a theoretical framework called the PAC learning model, and due to Valiant [19]. The question asked by boosting is whether *weak* learners that perform slightly better than random guessing can be *boosted* into a significantly more accurate strong learning algorithm. A breakthrough came in 1995 with the development of the Adaboost algorithm [20] , which iteratively determines weights on several weak learning algorithms to output a final strong learner.

The main idea of the algorithm is to maintain a distribution or set of weights over the training set. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.

Fig. 2 shows our application of AdaBoost in meta-classification. Table II outlines the actual steps used in the AdaBoost algorithm. In Table II, S is the training set of meta-features obtained from stage 1 of the classifier in Fig. 2, and N denotes the total number of training samples. The hypotheses $\{f_t(x)\}_{t=1}^{T}$ correspond to the weak learners which are Neural Network based classifiers trained on S, and T denotes the number of weak learners. In our design, typical values of T varied from $10 - 20$. The final hypotheses $F(x)$ is a weighted majority vote of the T weak hypotheses where β_t is the weight assigned to f_t.

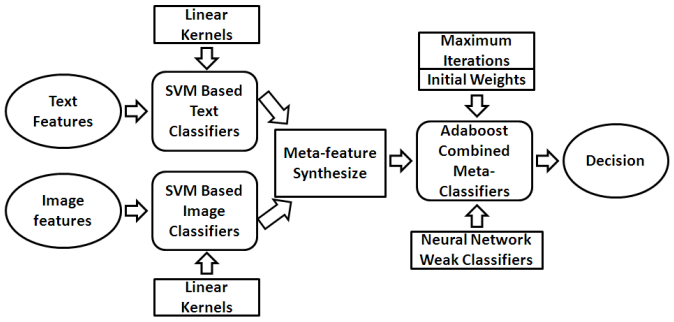

Fig. 2. Meta-Classification Via Adaboost

IV. EXPERIMENTAL RESULTS

Recall and *Precision* are widely used to evaluate the performance of classification algorithms [3].

Let $h(\mathbf{x})$ denote the designed classifier, m the number of classes, $\mathbf{x}_j$ be the j^{th} test sample, whose true label is $y_j \in \{1, 2, \ldots, m\}$. We first define f_{++} as the number of test data samples for which $h(\mathbf{x}_j) = k$ and $y_j = k$, f_{+-} as the number of test data samples for which $h(\mathbf{x}_j) = k$ and $y_j \neq k$, and

TABLE II
THE ADABOOST LEARNING ALGORITHM [21]

1. Given (x_i, y_i), $i = 1, 2, \ldots, N$ where $x_i \in S$, $y_i \in \{1, -1\}$
2. Initialize $D_1(i) = \frac{1}{N}$, $i = 1, 2, \ldots, N$
3. Repeat for $t = 1, 2, \ldots, T$:
   (a) Train weak learner using distribution D_t
   (b) Calculate weak hypothesis $f_t : S \longrightarrow \{1, -1\}$ with error ϵ_t
   (c) Choose $\beta_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$
   (d) Update $D_t(i)$ to $D_{t+1}(i)$ using β_t, y_i, and $f_t(x_i)$
3. Output decision function $F(x) = sign[\sum_{t=1}^{T} \beta_t f_t(x)]$.

f_{-+} as the number of test data samples for which $h(\mathbf{x}_j) \neq k$ and $y_j = k$.

$Recall$ and $Precision$ for the k^{th} class, where $k \in \{1, 2, \ldots, m\}$, are defined as:

$$Recall_k(h) = \frac{f_{++}}{f_{++} + f_{-+}} \tag{2}$$

$$Precision_k(h) = \frac{f_{++}}{f_{++} + f_{+-}}. \tag{3}$$

Clearly, there is a tradeoff between the two quantities. Their harmonic mean F_k is often used to evaluate the performance of the classifier:

$$F_k(h) = \frac{2Recall_k(h)Precision_k(h)}{Recall_k(h) + Precision_k(h)} \tag{4}$$

The higher F_k is, the better the classification performance.

A. Wikipedia Database

A key challenge in evaluating joint text and image feature based document classification is that unlike the problem of text based document classification, *realistic* benchmark databases are not readily available.

Results on synthetically generated documents are not convincing and can often be misleading. To work with real-world multimedia documents, we use webpage documents from the Wikipedia selection for schools 2008/2009 edition [22]. This database contains about 5500 articles and is about the size of a 20-volume encyclopedia. Each document contains a text article and in general varying number of images ranging from 3 to 20. As we conjectured earlier, the images in the documents can be qualitatively seen to relate well to the text in the document. For example, an article about clarinet [23] from the class labeled "music" contains images of clarinets and people playing clarinets. The statistics and details of the database are listed in Table III.

The merits of working with the Wikipedia database are that it is widely used, easily accessible, and also comes with pre-assigned manual labels. The significant challenge however is that many document categories contain a small number of documents and that means fewer documents are available for training.

We experimented with four data collections which are listed in Table IV. The first two sets are composed of three to four distinct categories. In realistic databases, sometimes categories can be a subset of main categories. Hence, we use two document collections (labeled 3 and 4) where the individual categories can be conceptually interpreted as subsets of an umbrella category.

TABLE III
CATEGORIES OF WIKIPEDIA SELECTION FOR SCHOOLS

Name	Number of Documents	Number of Images	Number of Subcategories
art	86	602	1
business	141	1273	4
citizenship	311	2415	9
design and technology	282	2655	5
everyday life	430	3882	11
geography	1198	19273	17
history	836	6388	11
information technology	86	354	5
language and literature	197	842	8
math	263	1027	1
music	168	654	4
people	751	5326	19
religion	176	1029	6
science	1205	7226	3

TABLE IV
SELECTED DATASETS

Set		Name	Number of Documents	Number of Images
1	1.1	art	86	602
	1.2	business	141	1273
	1.3	information technology	86	354
	1.4	religion	176	1029
2	2.1	design and technology	282	2655
	2.2	math	263	1027
	2.3	music	168	654
3	3.1	science.biology	771	4526
	3.2	science.chemistry	162	793
	3.3	science.physics	251	1907
4	4.1	music.musical genres styles eras and events	31	146
	4.2	music.musical instruments	37	220
	4.3	music.musical recordings and compositions	33	37
	4.4	music.performers and composers	67	251

B. Results

We randomly partition the 4 document collections into training and testing documents. On an average, only 10 documents were used in each category for training.

$Recall$, $Precision$ and their harmonic mean F are reported for each class in Table V for the document collection 1. Analogous results for document collections 2-4 are reported in Tables VI - VIII. In these tables, R_I, P_I represent $Recall$ and $Precision$ of document classification based on image features only. Likewise, R_T, P_T represent $Recall$ and $Precision$ of document classification based on text features only. The harmonic mean values F_I, F_T are analogous. Symbols with subscripts $ADA - i, i = 1, 2$ represent meta-classification results using Adaboost with image feature vector combination strategy i (as defined in Table I), and symbols with subscripts $SVM - i$ represent analogous results using the SVM based meta-classifier. $MAXIMP_i$ stands for maximum improvement for the corresponding image feature vector combination

strategy i. This quantity is given by

$$max(F_{ADA-i} - max(F_T, F_I), F_{SVM-i} - max(F_T, F_I)).$$

TABLE V
RECALL, PRECISION AND F FOR SET 1

	1.1	1.2	1.3	1.4	Average
R_I	51.7%	25.0%	37.4%	61.4%	43.9%
R_T	84.5%	98.6%	68.2%	69.3%	80.2%
R_{ADA-1}	96.6%	86.1%	76.6%	85.0%	86.1%
R_{ADA-2}	87.9%	97.2%	68.2%	78.0%	82.8%
R_{SVM-1}	91.4%	97.2%	66.4%	76.4%	82.8%
R_{SVM-2}	87.9%	97.2%	62.6%	80.3%	82.0%
P_I	41.1%	32.1%	50.0%	50.3%	43.4%
P_T	90.7%	48.3%	98.7%	98.9%	84.1%
P_{ADA-1}	86.2%	67.4%	96.5%	88.5%	84.6%
P_{ADA-2}	82.3%	63.6%	91.3%	88.4%	81.4%
P_{SVM-1}	80.3%	62.5%	92.2%	89.0%	81.0%
P_{SVM-2}	77.3%	63.1%	93.1%	88.7%	80.5%
F_I	45.8%	28.1%	42.8%	55.3%	43.0%
F_T	87.5%	64.8%	80.7%	81.5%	78.6%
F_{ADA-1}	91.1%	75.6%	85.4%	86.7%	84.7%
F_{ADA-2}	85.0%	76.9%	78.1%	82.8%	80.7%
F_{SVM-1}	85.5%	76.1%	77.2%	82.2%	80.2%
F_{SVM-2}	82.3%	76.5%	74.9%	84.3%	79.5%
$MAXIMP_1$	3.6%	11.2%	4.8%	5.3%	6.2%
$MAXIMP_2$	-2.5%	12.1%	-1.8%	1.4%	2.3%

TABLE VI
RECALL, PRECISION AND F FOR SET 2

	2.1	2.2	2.3	Average
R_I	91.7%	75.5%	43.9%	70.4%
R_T	65.6%	94.3%	69.9%	76.6%
R_{ADA-1}	93.7%	90.6%	78.9%	87.7%
R_{ADA-2}	91.3%	91.7%	73.2%	85.4%
R_{SVM-1}	92.1%	91.2%	82.9%	88.7%
R_{SVM-2}	92.5%	90.6%	78.1%	87.1%
P_I	71.6%	94.2%	60.0%	75.3%
P_T	95.4%	61.4%	86.9%	81.2%
P_{ADA-1}	88.1%	96.7%	81.5%	88.8%
P_{ADA-2}	86.8%	89.3%	85.7%	87.3%
P_{SVM-1}	90.3%	94.6%	81.6%	88.8%
P_{SVM-2}	87.6%	91.1%	87.3%	88.7%
F_I	80.4%	83.8%	50.7%	71.6%
F_T	77.7%	74.3%	77.5%	76.5%
F_{ADA-1}	90.8%	93.6%	80.2%	88.2%
F_{ADA-2}	89.0%	90.5%	78.9%	86.1%
F_{SVM-1}	91.2%	92.8%	82.3%	88.8%
F_{SVM-2}	90.0%	90.9%	82.4%	87.8%
$MAXIMP_1$	10.8%	9.7%	4.8%	8.4%
$MAXIMP_2$	8.6%	6.7%	3.5%	6.3%

The benefits of combining the two modalities are readily apparent from the results in Tables V-VIII. The average improvement in the harmonic mean, i.e., F, is also visualized in Fig. 3. The maximum improvement is usually around 7%, and the average F is above 80%.

Further, the improvements are seen for both SVM as well as AdaBoost based meta-classification. This validates our intuition that augmenting classifier decisions with image features can improve document classification results and does not acutely depend on the choice of the classifier. These

TABLE VII
RECALL, PRECISION AND F FOR SET 3

	3.1	3.2	3.3	Average
R_I	17.2%	58.9%	82.0%	52.7%
R_T	92.9%	61.0%	55.3%	69.7%
R_{ADA-1}	86.3%	92.9%	73.7%	84.3%
R_{ADA-2}	86.0%	90.8%	72.4%	83.0%
R_{SVM-1}	88.9%	78.7%	76.0%	81.2%
R_{SVM-2}	88.5%	76.6%	72.8%	79.3%
P_I	88.6%	57.2%	22.4%	56.1%
P_T	83.8%	65.2%	81.6%	76.9%
P_{ADA-1}	94.3%	60.6%	78.8%	77.9%
P_{ADA-2}	93.4%	59.5%	78.5%	77.1%
P_{SVM-1}	92.4%	70.3%	72.7%	78.4%
P_{SVM-2}	91.8%	67.5%	70.5%	76.6%
F_I	28.8%	58.0%	35.2%	40.7%
F_T	88.1%	63.0%	65.9%	72.4%
F_{ADA-1}	90.1%	73.4%	76.2%	79.9%
F_{ADA-2}	89.5%	71.9%	75.3%	78.9%
F_{SVM-1}	90.6%	74.2%	74.3%	79.7%
F_{SVM-2}	90.1%	71.8%	71.7%	77.8%
$MAXIMP_1$	2.5%	11.2%	10.3%	8.0%
$MAXIMP_2$	1.4%	8.9%	9.4%	6.6%

TABLE VIII
RECALL, PRECISION AND F FOR SET 4

	4.1	4.2	4.3	4.4	Average
R_I	15.4%	100.0%	40.0%	63.3%	54.7%
R_T	69.2%	92.6%	80.0%	67.3%	77.3%
R_{ADA-1}	53.8%	100.0%	80.0%	79.6%	78.4%
R_{ADA-2}	53.9%	100.0%	100.0%	77.6%	82.9%
R_{SVM-1}	69.2%	96.3%	40.0%	87.8%	73.3%
R_{SVM-2}	69.2%	96.3%	80.0%	83.7%	82.3%
P_I	50.0%	55.1%	33.3%	88.6%	56.8%
P_T	69.2%	71.5%	30.8%	100.0%	67.9%
P_{ADA-1}	87.5%	75.0%	40.0%	97.5%	75.0%
P_{ADA-2}	70.0%	73.0%	71.4%	95.0%	77.4%
P_{SVM-1}	75.0%	89.7%	50.0%	87.8%	75.6%
P_{SVM-2}	75.0%	81.3%	66.7%	93.2%	79.0%
F_I	23.5%	71.1%	36.4%	73.8%	51.2%
F_T	69.2%	80.7%	44.4%	80.5%	68.7%
F_{ADA-1}	66.7%	85.7%	53.3%	87.6%	73.3%
F_{ADA-2}	60.9%	84.4%	83.3%	85.4%	78.5%
F_{SVM-1}	72.0%	92.9%	44.4%	87.8%	74.3%
F_{SVM-2}	72.0%	88.1%	72.7%	88.2%	80.3%
$MAXIMP_1$	2.8%	12.2%	8.9%	7.3%	7.8%
$MAXIMP_2$	2.8%	3.7%	38.9%	8.6%	13.5%

improvements are particularly remarkable in view of the fact that only about 10 training documents were used to train the classifier.

Finally, two observations are worth making. First, the strategy to combine features from multiple images seems to make an impact based on which document collection is tested, i.e. one may be better than the other based on the categories there in. This is intuitively satisfying and hence prior knowledge may be used in picking one of the two strategies. Second, note that in the classification results based on individual modalities, text based document classification is almost always superior to document classification based on image features. This points towards an ability to improve results further by extracting more semantically based and high-level features from images. We

expect to address that in future work.

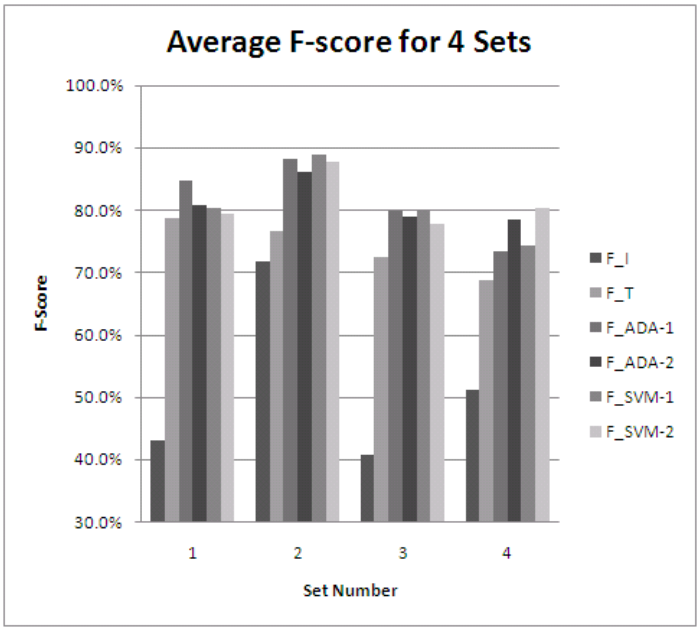

Fig. 3. F-score for 4 Sets

V. Conclusion and Future Work

This paper proposes a multi-modal signal processing approach to the problem of document classification. Inspired by analogous work in other domains, we design classifiers that can exploit the benefits of jointly operating on text as well as image features - the two key modalities of interest in modern documents. The primary contribution of this paper is in a meta-classification strategy that can effectively exploit the correlation between text and image data. Both support-vector machines and AdaBoost based meta-classifiers are designed and perform very well on real-world webpage document collections from Wikipedia, even as the number of training documents is severely limited. Experimental outcomes also sensitize us towards the need for better image feature extraction that can aid in document classification, and effectively complement text. Future work will also focus on statistical modeling of joint dependencies so as to enable better fusion of the two modalities.

References

[1] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1997, pp. 412–420.

[2] G. Salton, "Developments in automatic text retrieval," *Science*, 1991.

[3] T. Joachims, *Learning to classify text using support vector machines*. Kluwer Academic Publishers, 2001.

[4] R. E. Mayer and R. Moreno, "Aids to computer-based multimedia and learning," *Learning and Instruction*, vol. 2, pp. 107–119, 2002.

[5] Y. Wang, Z. Liu, and J. C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, 2000.

[6] W.-H. Lin and A. Hauptmann, "News video classification using svm-based multimodal classifiers and combination strategies," in *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*. ACM, 2002, pp. 323–326.

[7] T. D. A. Quattoni and M. Collins, "Learning visual representations using images with captions," *Computer Vision Pattern Recognition*, 2007.

[8] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts," *Multimedia*, pp. 112–121, 2005.

[9] W.-H. Lin and A. Hauptmann, "A meta-classification of multimedia classifiers," *Int. Workshop on Knowledge Discovery in Multimedia and Complex Data*, 2002.

[10] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *European Conference on Computational Learning Theory*, 1995, pp. 23–37.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *European Conference on Computer Vision*. Springer, 2006. [Online]. Available: http://lear.inrialpes.fr/pubs/2006/NJT06

[13] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," June 2006, pp. 13–13.

[14] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[15] V. N. Vapnik, *The Nature of Statistical Learning Thoery*. New York, NY, USA: Springer, 1995.

[16] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, pp. 121–167, 1998.

[17] S. Abe, *Support Vector Machines for Pattern Classification*. London, UK: Springer, 2005.

[18] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[19] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, pp. 1134–1142, 1984.

[20] Y. Freund and R. E. Schapire, "A short introduction to boosting," *Journal of Artificial Intelligence*, pp. 771–780, 1999.

[21] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of Statistics*, vol. 28, 2000.

[22] (2008, Oct.) 2008/9 wikipedia selection for schools. [Online]. Available: http://schools-wikipedia.org

[23] (2008, Oct.) Clarinet. [Online]. Available: http://schools-wikipedia.org/wp/c/Clarinet.htm