This paper describes work performed at the University of Sunderland as part of the EU-funded VITALAS project. Text feature vectors, extracted from the TRECVID video data set, were submitted to an SVM-Light implementation of Support Vector Machine, which aimed to label each video shot with the relevant concepts from the 101-concept MediaMill set. Sunderland also developed a search engine designed to match text queries derived from the test data against concept descriptors derived from the training data using the TF.IDF measure. The search engine-based approach outperformed SVM-Light, but did not perform overall as well as the MediaMill baseline for text feature extraction. However, the search-engine approach is much simpler than the supervised learning approach of MediaMill, and did outperform the MediaMill baseline for 31 of the 101 concept categories.

1. INTRODUCTION

The experiments described in this paper were undertaken as part of the EU-funded Video and Image Indexing and Retrieval in the Large Scale (VITALAS) project, which aims to provide a prototype system dedicated to the intelligent access to professional multimedia archives. The VITALAS system will initially be developed as a B2B tool, but the intention is to develop technologies which can also be used by the wider public with multimedia content search engines. One of the main objectives of the project is to investigate novel approaches to cross-media indexing, with the aim of scaling beyond the 200 to 300 concepts handled by current state-of-the-art systems. The experiments reported here are concerned purely with monomedia text features, while the project as a whole is concerned with multimedia.

The data set employed in our experiments was obtained from the 2005 TREC Video Retrieval Evaluation (TRECVID) data set [13] and the MediaMill Semantic Video Search Engine framework [14]. TRECVID is an international benchmarking activity, which provides a large test collection and means for researchers interested in video information retrieval to compare their results. TRECVID is co-sponsored by the National Institute of Standards and Technology (NIST) and other U.S. government agencies. Since 2003, TRECVID data has consisted of broadcast news. The 2005 development set contains 85 hours of training data and another 85 hours of test data, comprising a total of 43,907 shots; 30,093 for training and 13,814 for testing. In 2003-2004 the speech accompanying the video data was entirely in English, but in 2005, Chinese and Arabic broadcasts were also included.

MediaMill provides a lexicon of 101 semantic concepts, and ground truth data with the most suitable concept, or concepts, for each video shot. Also, it provides English-language transcripts of the text of the 2005 English TRECVID broadcast news, and machine translation transcripts of the Chinese and Arabic TRECVID broadcasts for the same year. It should be observed that the translated texts contain a high number of errors. MediaMill allows the comparison of methods for each of the intermediate steps in video indexing. This is achieved by decomposing the overall video indexing problem into a number of experiments, each of them examining an individual component or combination of components. The MediaMill experiments of greatest relevance to the work described in this paper are as follows:

Experiment 2, which focuses on text analysis: The goal is to learn a list of video shots, ranked according to their relevance to each of the 101 concepts, given a vector of extracted text features.

Experiment 5, which focuses on both text and content-based video analysis: The goal is to learn the same ranked list as in experiment 2, but this time the list is the best list—as measured by mean average precision (MAP)—produced out of a series of experiments, involving monomedia text, monomedia visual features, and text and visual features combined by early and late fusion.

Snoek et al. [14] offer a detailed description of the two experiments mentioned above and the way they carried them out. Since the focus of this paper is on text, we compare our two approaches with the text analysis baseline of experiment 2. Experiment 5 allows us to get a feel for how much of the optimal multimedia performance is achievable by text analysis alone, and how much better the performance can be with the integration of visual data.
2. SVM-LIGHT

Support vector machines have been found to be very effective in various categorisation tasks [7, 9, 16, 18]. Support vector machines are based on the structural risk minimisation principle for which error-bound analysis has been motivated theoretically [17]. The idea of SVM is to find an optimal decision surface that separates the positive training documents from negative ones over a vector space. Linear SVM has shown its effectiveness in many cases [5, 7, 10, 19]. A decision surface in a linearly separable vector space is a hyperplane, and a linear SVM aims to find the hyperplane that maximises the margin distances between the hyperplane and the support vectors. More detailed descriptions of SVM can be found in [6, 12].

In our study, we compared a linear inductive SVM, which aims to find the hyperplane that maximises the margin distances between the hyperplane and the support vectors, with a linear transductive SVM, which finds the hyperplane based on both the training and the test data and thus is more robust when fewer training data is available [8]. We employ the freely-available SVM-Light software offered at http://svmlight.joachims.org/. Both the inductive and transductive SVM were trained using the one-against-all method, and we fixed the trade-off between training error and margin \( \epsilon = 1 \) and kept the cost-factor \( j = 1 \) at all times.

Following Adams et al. [1], the automatic speech recognition (ASR) transcriptions provided by MediaMill for each shot were transformed into text feature vectors, which were the input to SVM-Light. After tokenisation—breaking the text into individual words—, removal of stop words—very frequent words with low information content—and the use of stemming rules—we used Porter’s stemming rules [11]—to reduce all grammatical variants of a word to a common stem, each element of each vector corresponded to the number of times a particular word stem was encountered in a shot.

Prior to textual feature extraction, we stretched the shot boundaries. That is to say, we associated with each shot not only its own text, but that assigned to the immediately previous and following shots. This is due to the fact that names, as well as other indicative words, are often mentioned in broadcast news just before or after a picture of a person is visible [15].

We omitted the shots from the training data and the test data that either did not contain any caption after stemming and stop-word removal or did not have any concept associated with them, because they did not provide the classifier any information. The resulting training data contained 24,846 shots and the resulting test data contained 9,646 shots.

3. A SEARCH ENGINE-BASED APPROACH

In the field of information retrieval, which underpins search engine technology, one task is to find the relatively small set of documents—typically, Web pages—which are of interest to someone from an electronic repository containing a much larger set of documents—typically, the entire Web—, most of them not immediately relevant to the information seeker. In an interaction with a search engine such as Google, the information seeker inputs a set of words reflecting their topic of interest—referred to as the query—, and the engine compares these words against the index terms which have been chosen to best represent the content of each document. Documents whose index terms best match the query terms are shown to the information seeker, in order of the degree of match. Although document indexing can be done manually, in large collections such as the Web this must be done automatically. A number of statistical measures exist for this: the simplest one is raw frequency, but the most widely used is called TF.IDF.

The principle behind the raw frequency approach is that the words which tell us most about the content of a document are the mid-frequency words within that document. Highly frequent words, such as the, a, and of, are much less useful and are routinely disregarded, if they occur on standard lists of such words, like the one created for the SMART information retrieval system [3]. Terms which occur only once in a document, such as the new in the previous paragraph, also tell us little about the content of the document. An arbitrary lower bound must be chosen, such that words found less than this number of times are excluded from the index.

The main weakness with the raw frequency approach is that we need to take into account not only the frequency of a word in a document—as we do—, but also how often the word is found in other documents.

The more sophisticated TF.IDF measure does take into account both the raw frequency of the word in a particular document (TF, or term frequency), and the inverse of the number of documents in the collection in which the word appears (IDF, or inverse document frequency).

In this section we will examine the use of the TF.IDF measure, not for conventional document indexing, but for the related task of working out which words in the transcribed text accompanying the MediaMill training set are most typical of one of the 101 MediaMill concepts, as opposed to any of the other 100 concepts.

The basic formula for TF.IDF that we used in our study is as follows [2]:

\[
W_{kd} = f_{kd} \cdot \log \left( \frac{N_{Doc}}{D_k} \right),
\]
where \( w_{kd} \) is the weight reflecting the typicality of term \( k \) with respect to shot \( d \), \( f_{kd} \) is the raw frequency of term \( k \) in shot \( d \), \( NDoc \) is the total number of shots in the collection, and \( D_k \) is the number of shots which contain term \( k \) at least once. The highest TF.IDF scores are given to those terms which are common in the shot we are looking at, but do not occur in many other shots. Although not used here, variants of TF.IDF exist to take into account such factors as differing document lengths.

When indexing in an information retrieval context, we are interested in the vocabulary characteristic of a document, such as a Web page, but here we are interested in the vocabulary characteristic of a MediaMill concept, in order to index that concept. Also, rather than matching our document index terms against a user-selected query, here our “queries” are the text associated with individual video shots in the MediaMill test set. A worked example is shown in Table 1.

<table>
<thead>
<tr>
<th>Video Shot</th>
<th>Transcribed Text</th>
<th>Ground Truth Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>benazir bhutto bus bomb</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>pakistan karachi</td>
<td>2, 3, 55, 96</td>
</tr>
<tr>
<td>3</td>
<td>bhutto pakistan people party</td>
<td>1, 4, 101</td>
</tr>
<tr>
<td>4</td>
<td>poster flag benazir bhutto</td>
<td>2, 17, 98</td>
</tr>
</tbody>
</table>

Table 1: Worked example for TF.IDF

To find the degree of relevance of the word bhutto, found in the training set, to concept 1, we first combine the vocabulary of all the shots declared by the ground truth data as being relevant to that concept—i.e., shots 1 and 3. The combined vocabulary is benazir, bhutto—twice—, bus, bomb, pakistan, people, party. The term frequency of bhutto is thus 2. Altogether there are 101 MediaMill concepts, so \( NDoc \) is 101. \( D_k \) is the number of concepts which corresponds to shots containing the term bhutto at least once, which is 6—concepts 1, 2, 4, 17, 98 and 101. Thus, the weight reflecting the typicality of the word bhutto with respect to concept 1 is

\[
2 \cdot \log \left( \frac{101}{6} \right) = 1.226.
\]

To illustrate that the technique can sometimes be very effective, the seven highest scoring (stemmed) words by TF.IDF with respect to the concept allawi were iraqi (30.2), govern (17.6), prime (16.4), minist (15.7), militari (13.9), iraq (13.2) and arm (8.8). A subject appraisal of these words suggests that as a set they are indeed related to allawi. The highest scoring words with respect to clinton were lewinski (12.2), clinton (7.7), monica (5.9), cabin (3.6), peac (3.5), david (3.3), and presid (3.1).

One popular formula for matching a set of query terms against a set of document index terms is the Cosine Similarity Coefficient [2], which is given by the formula:

\[
\text{Sim}(D_i, Q_j) = \frac{\sum_{k=1}^{t} (D_{Term_{ik}} \cdot Q_{Term_{jk}})}{\sqrt{\sum_{k=1}^{t} (D_{Term_{ik}})^2 \cdot \sum_{k=1}^{t} (Q_{Term_{jk}})^2}}.
\]

One interpretation of this formula is that \( D_{Term_{ik}} \) is the TF.IDF weight of index term \( k \) in document \( i \), \( Q_{Term_{jk}} \) is the number of times term \( k \) appears in query \( j \), \( t \) is the total number of terms in the vocabulary of the system, and \( \text{Sim}(D_i, Q_j) \) is the similarity between the document and the query in the range 0—no overlap at all between the query
and document terms—and 1—the query terms and the
document terms are identical.

In our experiments, the cosine similarity measure was
used to determine which MediaMill concepts were most
relevant to the transcribed text from each of the video shots
in the training set. The text for each shot was stop-listed and
stemmed, and became the “query” terms. The sets of words
determined previously by the TF.IDF method as being most
typical of each MediaMill concepts were regarded as
“document” terms. The concepts were ranked with respect
to each “query”, according to how well the “query” terms
matched the concept-related terms in each case.

4. RESULTS

The performance metric used for our experiments was the
mean average precision (MAP) achieved for each concept
category [14]. The results, for each of the 101 MediaMill
concept categories, are shown in Figure 1. Only the linear
inductive SVM is plotted in Figure 2. Although the
differences in MAP for the two forms of SVM were very
slight, the inductive one outperformed the transductive one
for 48 categories, while the transductive SVM performed
better for only 28 concepts. Both forms performed equally
well for the other 25 categories. Using the Wilcoxon
matched pairs signed ranks test, the inductive SVM
performed significantly better ($z = 4.99, p < 0.001$). The
search engine approach gave better MAP than the SVM for
80 of the concepts, while the SVM performed better for 18.
According to Wilcoxon test, this meant that the search
engine approach performed significantly better than the
SVM ($z = 4.99, p < 0.001$). The search engine approach
performed less well than the MediaMill experiment 2
baseline for 66 categories, but better for 31 categories. The
Wilcoxon test showed that the experiment 2 baseline was
significantly better than the search engine approach ($z = 4.84, p < 0.001$). However, we feel that reasonable results
had been obtained by using the much simpler, novel search
engine approach introduced in this paper.
6. CONCLUSION

In this paper, we have compared the effectiveness of a linear inductive SVM, a linear transductive SVM and a novel search-engine based approach for concept selection against two MediaMill baselines, one for their best approach based on text alone and another one for their best multimedia approach. All the techniques described here have the advantage of language independence. The linear inductive SVM outperformed the linear transductive SVM, but was in turn outperformed by the novel search engine approach. The search engine approach has the advantage of speed, taking 68 minutes to build the index compared with about 2.5 hours to train the inductive SVM and about 26 hours to train the transductive SVM. However, it did not perform as well as either MediaMill baseline. One reason for this difference may be that our search engine did not use stretched shots. The MediaMill experiments used a different type of SVM to SVM Light called LIBSVM with a radial basis function. A Wilcoxon sign test showed that the multimedia baseline outperformed the monomedia baseline ($z = 8.39, p < 0.001$) showing that valuable additional information is obtained from the visual features over and above that obtainable from text alone. In fact the search engine approach described here could be extended to incorporate visual features. Visual features have been extracted by CERTH-ITI (http://www.iti.gr/) for the ground truth test set, so we could learn which of these video features correspond with which concepts using the TF.IDF measure. This would enable the training data images to be labelled with concepts matching their extracted visual features as well as those matching the text of their captions.

7. ACKNOWLEDGMENTS

This research was supported under the EU-funded Video and Image Indexing and Retrieval in the Large Scale (VITALAS) project (project number FP6-045389). The authors are very grateful to CERTH-ITI (Greece) and the Fraunhofer Institute (Germany) for closely working with the University of Sunderland in the multimedia efforts of the VITALAS project.

8. REFERENCES


