Joint Visual-Text Modeling for Automatic Retrieval of Multimedia Documents

G. Iyengar
IBM TJ Watson Research Center
giyengar@us.ibm.com

P. Duygulu
Bilkent University
duygulu@cs.bilkent.edu.tr

S. Feng
University of Massachusetts, Amherst
slfeng@cs.umass.edu

P. Ircing
Univ. West Bohemia
ircing@kky.zcu.cz

S. P. Khudanpur
Johns Hopkins University
khudanpur@jhu.edu

D. Klakow
Saarland Univesity
dietrich.klakow@lsv.uni-saarland.de

M. R. Krause
Georgetown University
mrk6@georgetown.edu

R. Manmatha
University of Massachusetts, Amherst
manmatha@cs.umass.edu

H. J. Nock
IBM TJ Watson Research Center
hjnwork@hotmail.com

D. Petkova
Mt. Holyoke College
dipetkov@mtholyoke.edu

B. Pytlik
Johns Hopkins University
bep@cs.jhu.edu

P. Virga
Johns Hopkins University
pvirga@clsp.jhu.edu

ABSTRACT

In this paper we describe a novel approach for jointly modeling the text and the visual components of multimedia documents for the purpose of information retrieval(IR). We propose a novel framework where individual components are developed to model different relationships between documents and queries and then combined into a joint retrieval framework. In the state-of-the-art systems, a late combination between two independent systems, one analyzing just the text part of such documents, and the other analyzing the visual part without leveraging any knowledge acquired in the text processing, is the norm. Such systems rarely exceed the performance of any single modality (i.e. text or video) in information retrieval tasks. Our experiments indicate that allowing a rich interaction between the modalities results in significant improvement in performance over any single modality. We demonstrate these results using the TRECVID03 corpus, which comprises 120 hours of broadcast news videos. Our results demonstrate over 14% improvement in IR performance over the best reported text-only baseline and ranks amongst the best results reported on this corpus.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.4.9 [**Image Processing and Computer Vision**]: Applications

General Terms

Retrieval Models, Experimentation

Keywords

Joint Visual-Text Models, TRECVID, Multimedia Retrieval Models

1. INTRODUCTION

There has been a renewed spurt of research activity in Multimedia Information Retrieval. This can be partly attributed to the emergence of a NIST-sponsored video analysis track, namely TRECVID[16], coinciding with a renewed interest from industry and government in developing techniques for mining vastly growing quantities of multimedia data.

Most of the state-of-the-art multimedia retrieval systems are either pure text-based retrieval systems or at best a late fusion of speech-based retrieval techniques and image content-based retrieval techniques. It is our hypothesis that such system-level integration allows only limited exploitation of cues that occur in the different modalities. In addition, techniques used in retrieval systems using images and speech differ vastly and this further inhibits interaction between these systems for multimedia information retrieval. For instance, if the query words have been incorrectly recognized then speech-based retrieval systems may fail. Current systems back-off to image content-based searches and since image retrieval systems perform poorly for finding images related by semantics, the overall performance of such late-fusion systems can be poor. This situation is exacerbated in cross-lingual information retrieval where machine translation can further degrade the text transcript. Previous results on the TRECVID corpus indicate that late-fusion systems achieve little gain, if any, in retrieval performance over unimodal systems (cf. systems participating at TRECVID2003 benchmark [16]). Further, the dominant retrieval paradigms in TRECVID are *manual* (human-in-the-loop to process the query once) and *interactive* (human-in-the-loop to process the query and provide feedback to the retrieval system) which complicate detailed analysis of such systems. In this paper, we build automatic retrieval systems (no human intervention or interaction) to isolate the algorithmic issues from user interface design issues. This is a relative novelty as well as there are only a handful of systems that are fully automatic[1].

In this paper, we investigate a unified approach to multimedia information retrieval. We represent a multimedia document in terms of visual and textual tokens to build various joint statistical models. This allows us to treat multimedia retrieval as a generalized version of statistical text retrieval—one where we retrieve documents made up of words and visual tokens. With joint visual-text modeling, we demonstrate that we can better represent the relationships between words and the associated visual cues. In this work, we phrase the multimedia retrieval task in terms of a *generative* model. That is, we model the different ways the query q is generated from the document d. We then rank the documents using $p(\mathbf{d}|\mathbf{q})$, or given a query $\mathbf{q}$, the probability that the document $\mathbf{d}$ generated it. To illustrate and validate the usefulness of this approach, we build automatic multimedia retrieval systems, and present experimental results on the TRECVID03 corpus and queries.

2. RELATED WORK

The dominant approach in state-of-the-art systems for ad-hoc video retrieval is to perform a text-based search on the speech recognition transcript associated with the video data[16, 17]. Pure visual-only systems generally have performed very poorly in comparison with text-based systems on this task. Whilst some groups have shown multimodal systems that outperform some text-only systems, the best performing systems have been text-based[16, 17]. In addition, these systems make the assumption that there is a human-in-the-loop who interprets the statement of information need and interacts with the result set, refining the query and providing feedback. So, the emphasis is balanced between good initial retrieval and a well-designed user-interface. The lack of interaction between the text-based and video-based IR components can be mitigated by manually choosing appropriate means to combine the results. For instance, a query to locate a specific person when presented to the text-based system will retrieve segments of video where the person is mentioned (e.g. Bill Clinton, the President etc) but not necessarily shots containing his face. A subsequent interaction by the human in the loop refines the returned segments and eliminating those that do not contain faces (e.g. with the help of a face detector) can produce the desired result set. For example, see the system descriptions at TRECVID03[9, 24].

In this paper, we are interested in modeling the interaction between the speech transcript and the images that comprise a broadcast video segment. With this goal, we build automatic retrieval systems and shift the focus onto aspects of modeling the different parts of a multimedia document and query without considering the user interface issues. A similar approach was attempted by Westerveld and de Vries[22]. We show that their approach of modeling the query using multiple visual examples can be seen as a particular instance of our general framework. In particular, they model two of the four components that we detail in our framework. They model the relationship between the query text and document text and the relationship between the query visuals and document visuals. In our paper, we additionally model the cross-relationships between the query images and document text and viceversa.

[1]TRECVID 2005 is the first year with a recognized automatic system track.

3. RETRIEVAL MODELS

Given a query, $\mathbf{q}$, we want to rank documents, $\mathbf{d}$, according to $p(\mathbf{d}|\mathbf{q})$. Let us represent the visual part of a multimedia document by $\mathbf{d_v}$ and the textual part by $\mathbf{d_w}$, and similarly for the multimedia query $\mathbf{q}$. This can be expanded as below.

$$\begin{aligned} p(\mathbf{d}|\mathbf{q}) &= p(\mathbf{d_w}, \mathbf{d_v}|\mathbf{q_w}, \mathbf{q_v}) \\ &= \frac{p(\mathbf{q_w}, \mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v})p(\mathbf{d_w}, \mathbf{d_v})}{p(\mathbf{q_w}, \mathbf{q_v})} \end{aligned} \tag{1}$$

In Eq. 1 the denominator can be ignored for ranking documents given any query. In addition, we will assume that all documents are equally likely. Any relaxation of this assumption can be done externally and applied to all the models that we develop here. This simplifies Eq. 1 to

$$p(\mathbf{d}|\mathbf{q}) \propto p(\mathbf{q_w}, \mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) \tag{2}$$

There may not be enough data to jointly model the above, necessitating further simplifying assumptions. Eq. 2 will get factored into different forms depending on the modeling assumptions made. We begin by assuming that the query word tokens and visual tokens (`visterms`) are conditionally independent given the document. That is the right-hand side of Eq. 2 can be written down as

$$p(\mathbf{q_w}, \mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) = p(\mathbf{q_w}|\mathbf{d_w}, \mathbf{d_v}) \times p(\mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) \tag{3}$$

3.1 Linear Mixture Model

Consider the term $p(\mathbf{q_w}|\mathbf{d_w}, \mathbf{d_v})$. We can choose to approximate it with a linear mixture model, to further simplify the modeling task:

$$p(\mathbf{q_w}|\mathbf{d_w}, \mathbf{d_v}) \approx \lambda_w p(\mathbf{q_w}|\mathbf{d_w}) + (1-\lambda_w)p(\mathbf{q_w}|\mathbf{d_v}) \tag{4}$$

Now, each of the two sub-components can be independently estimated using two different models. Another choice is to ignore the second term (equivalent to setting the mixture weight $\lambda_w = 1$). We can model the visual term $p(\mathbf{q_v}|\mathbf{d_w}\mathbf{d_v})$ similarly:

$$p(\mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) \approx \lambda_v p(\mathbf{q_v}|\mathbf{d_w}) + (1-\lambda_v)p(\mathbf{q_v}|\mathbf{d_v}) \tag{5}$$

Putting it all together, we get

$$\begin{aligned} p(\mathbf{q}|\mathbf{d}) \approx\ & (\lambda_w p(\mathbf{q_w}|\mathbf{d_w}) + (1-\lambda_w)p(\mathbf{q_w}|\mathbf{d_v})) \times \\ & (\lambda_v p(\mathbf{q_v}|\mathbf{d_w}) + (1-\lambda_v)p(\mathbf{q_v}|\mathbf{d_v})) \end{aligned} \tag{6}$$

3.2 Log Linear Model

Below is a maximum-entropy inspired approach which is an alternative to the linear model. We will start with the problem of estimating

$$p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) \tag{7}$$

The full probability is difficult to estimate because of a lack of training data. Hence, we assume that only pair distributions (e.g. $p(\mathbf{d_w}, \mathbf{d_v})$ or $p(\mathbf{d_w}, \mathbf{q_v})$) can be reliably estimated. This amounts to a set of constraint equations:

$$\sum_{d_w, d_v} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{q_w}, \mathbf{q_v}) \quad (8)$$

$$\sum_{d_w, q_w} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{d_v}, \mathbf{q_v}) \quad (9)$$

$$\sum_{d_w, q_v} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{d_v}, \mathbf{q_w}) \quad (10)$$

$$\sum_{d_v, q_w} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{d_w}, \mathbf{q_v}) \quad (11)$$

$$\sum_{d_v, q_v} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{d_w}, \mathbf{q_w}) \quad (12)$$

$$\sum_{q_w, q_v} p(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) = p(\mathbf{d_w}, \mathbf{d_v}) \quad (13)$$

Using a maximum entropy approach a probability distribution can be found that satisfies all six constraints. Instead of doing a full maximum entropy approach, we will just do one iteration of the Generalized Iterative Scaling algorithm (GIS)[4].

Assuming statistical independence of all four random variables the initial distribution is:

$$p_0(\mathbf{d_w}, \mathbf{d_v q_w}, \mathbf{q_v}) = p(\mathbf{d_w})p(\mathbf{d_v})p(\mathbf{q_w})p(\mathbf{q_v}) \quad (14)$$

After one iteration of GIS (using shortform $p_1(\mathbf{d}, \mathbf{q})$ to represent $p_1(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v})$) we arrive at:

$$p_1(\mathbf{d}, \mathbf{q}) = \frac{1}{Z} p_0(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) \left(\frac{p(\mathbf{q_w}, \mathbf{q_v})}{p(\mathbf{q_w})p(\mathbf{q_v})}\right)^{\lambda_1}$$
$$\left(\frac{p(\mathbf{d_v}, \mathbf{q_v})}{p(\mathbf{d_v})p(\mathbf{q_v})}\right)^{\lambda_2} \left(\frac{p(\mathbf{d_v}, \mathbf{q_w})}{p(\mathbf{d_v})p(\mathbf{q_w})}\right)^{\lambda_3} \left(\frac{p(\mathbf{d_w}, \mathbf{q_v})}{p(\mathbf{d_w})p(\mathbf{q_v})}\right)^{\lambda_4}$$
$$\left(\frac{p(\mathbf{d_w}, \mathbf{q_w})}{p(\mathbf{d_w})p(\mathbf{q_w})}\right)^{\lambda_5} \left(\frac{p(\mathbf{d_w}, \mathbf{d_v})}{p(\mathbf{d_w})p(\mathbf{d_v})}\right)^{\lambda_6} \quad (15)$$

where Z is a normalization and the λ_i are weights for the six constraint equations. Note that the above is the standard exponential form of a MaxEnt model. Ignoring all terms that do not matter for the decision and also assuming a uniform distribution for $p(\mathbf{d_w}, \mathbf{d_v})$ gives:

$$p_1(\mathbf{d_w}, \mathbf{d_v}, \mathbf{q_w}, \mathbf{q_v}) \propto (p(\mathbf{d_v}, \mathbf{q_v}))^{\lambda_2} (p(\mathbf{d_v}, \mathbf{q_w}))^{\lambda_3}$$
$$(p(\mathbf{d_w}, \mathbf{q_v}))^{\lambda_4} (p(\mathbf{d_w}, \mathbf{q_w}))^{\lambda_5} \quad (16)$$

This can be transformed into

$$p_1(\mathbf{q_w}, \mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) \propto (p(\mathbf{q_v}|\mathbf{d_v}))^{\lambda_2} (p(\mathbf{q_w}|\mathbf{d_v}))^{\lambda_3}$$
$$(p(\mathbf{q_v}|\mathbf{d_w}))^{\lambda_4} (p(\mathbf{q_w}|\mathbf{d_w}))^{\lambda_5} \quad (17)$$

This framework has been tested in language modeling components of automatic speech recognition systems where it usually outperformed linear interpolation[12]. Note that it has only one more free parameter compared to the linear mixture model in its complete form, since one of the exponents can be set to one without any influence on the ranking of the documents. We note here that this approach uses the same component conditional probabilities as in Eq. 6. The model proposed by Westerveld and de Vries[22] can be seen as a special case of this model. In particular, they model the $p(\mathbf{q_v}|\mathbf{d_v})$ and $p(\mathbf{q_w}|\mathbf{d_w})$ components of the log-linear model. Their $p(\mathbf{q_w}|\mathbf{d_w})$ is a text-only system similar to the one we used in this paper.

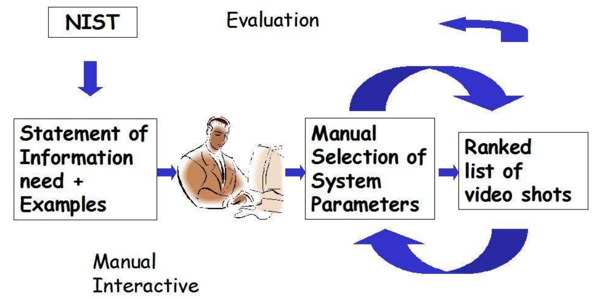

Figure 1: The manual and interactive system designs permitted by NIST in TRECVID evaluations

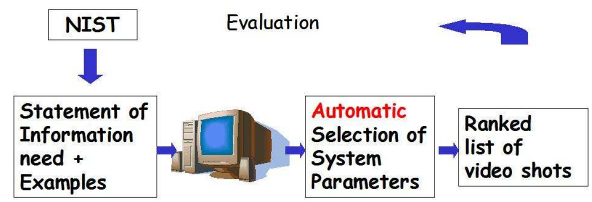

Figure 2: Automatic Multimedia Information Retrieval: System design

4. BASELINE SYSTEM

In multimedia retrieval tasks, text based systems have outperformed image content based systems by a wide margin. Therefore, we will compare the fusion systems with text based systems. In addition, while NIST permits *manual* and *interactive* query runs, we will restrict our experiments to *automatic* systems where there is no human intervention. This choice is to restrict our system design to only the algorithmic issues and ignore the user-interface issues. Figures 1 and 2 illustrate the differences between NIST and our system designs.

In the framework that we propose, the baseline system is obtained by setting $\lambda_w = 1$ and leaving out the visual component. We further assume that all the words in the document are independent of each other given the document, i.e. the bag of words document model. This results in a simple unigram language model over the words in a document[18]. We get

$$p(\mathbf{q_w}, \mathbf{q_v}|\mathbf{d_w}, \mathbf{d_v}) = p(\mathbf{q_w}|\mathbf{d_w}) = \prod_{i=1}^{m} p(q_{w_i}|\mathbf{d_w}) \quad (18)$$

where $\mathbf{q_{wi}}$ are the words in the query. $p(\mathbf{w}|\mathbf{d})$ can be modeled using a variety of smoothing techniques. For illustration, we use the Jelinek-Mercer smoothing to give us

$$p(\mathbf{w}|\mathbf{d}) = \alpha \frac{\#(\mathbf{w}, \mathbf{d})}{|\mathbf{d}|} + (\mathbf{1} - \alpha)\mathbf{p}(\mathbf{w}|\mathbf{C}) \quad (19)$$

where $\#(w, d)$ is the number of times the word w occurs in document d and $|d|$ is the total number of words in that document. C is the entire corpus of documents. In addition, we can attempt to relate query words to document words by performing semantic smoothing using a markov chain or estimating a stochastic dictionary using machine translation (see [13, 1] for examples of both approaches). For our baseline, we chose unigram modeling and smoothing with Dirichlet prior as this gave the best results on the test data. We also tried Jelinek-Mercer smoothing. However, it resulted in a 5-10% relative degradation in retrieval performance compared to using Dirichlet priors.

5. RELATING THE WORDS AND THE VISUAL PARTS OF THE DOCUMENT AND QUERY

One possible approach to joint visual-text retrieval is to build a direct model that relates words to parts of a picture. However, given the present state of computer vision, this is not a feasible task. Fortunately, the TRECVID data has been annotated with semantic concepts that cover essential parts of the pictures. This annotation is the result of the common annotation forum effort organized by NIST[15]. This annotation set consists of over 100 concepts manually marked on the 2003 development dataset. We selected a subset of 75 concepts that have more than 20 training examples in the development set for the purposes of this paper. Our approach is to build models from these concept annotations, and utilize these to relate the visual and textual parts of multimedia documents. We note that this approach is equivalent to introducing a hidden *Information Bottleneck* layer into the modeling framework. For further details on the Information Bottleneck method, see citation[21].

5.1 Single model with concept layer

In the following, the semantic concepts will be denoted by $\mathbf{c}$. In the previous sections $p(\mathbf{d}|\mathbf{q})$ has been decomposed into four different terms for textual and visual queries and the textual and visual parts of the documents.

One of the four terms is $p(\mathbf{q_w}|\mathbf{d_v})$, which we will discuss first. To use the concepts, a hidden layer is introduced as an information bottleneck; The probability is decomposed using the definition of conditional probabilities and finally, an independence assumption is made:

$$p(\mathbf{q_w}|\mathbf{d_v}) = \sum_{\mathbf{c}} p(\mathbf{q_w c}|\mathbf{d_v}) \quad (20)$$

$$= \sum_{\mathbf{c}} p(\mathbf{q_w}|\mathbf{c}\,\mathbf{d_v})p(\mathbf{c}|\mathbf{d_v}) \quad (21)$$

$$\approx \sum_{\mathbf{c}} p(\mathbf{q_w}|\mathbf{c})p(\mathbf{c}|\mathbf{d_v}) \quad (22)$$

- $p(\mathbf{c}|\mathbf{d_v})$ is one of the models trained as before. Only difference is that concept labels are used instead of the words. This describes the concept annotation detailed in Section 5.2.
- $p(\mathbf{q_w}|\mathbf{c})$ can be derived either from one of the previously trained document-concept models (e.g. $p(\mathbf{c}|\mathbf{d_w})$) or can be estimated independently. In practise, we have found that an adaptation step is desired to adapt $p(\mathbf{c}|\mathbf{d_w})$ for more suitable modeling of the query statistics.
- Approximating $p(\mathbf{q_w}|\mathbf{c},\mathbf{d_v})$ by $p(\mathbf{q_w}|\mathbf{c})$ is very crude. If the query is "Alan Greenspan" the concept will be "face" and such a model alone (even if perfect) will then return only faces.
- The same line of reasoning can be applied to the other three components $p(\mathbf{q_w}|\mathbf{d_w})$, $p(\mathbf{q_v}|\mathbf{d_w})$ and $p(\mathbf{q_v}|\mathbf{d_v})$. In each case, the result is a combination of model types already discussed earlier in the paper.

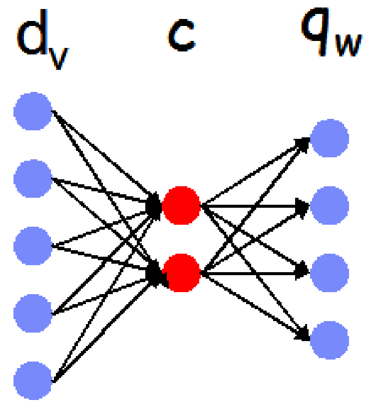

Figure 3: Illustration of the Information Bottleneck

Figure 3 illustrates the information bottleneck relating the document `visterms` and the query word tokens.

5.2 Relating Query Words with Document Visuals

One approach for estimating the probability of the concepts given the visual features of a keyframe ($p(\mathbf{c}|\mathbf{d_v})$) is to learn the correspondences between concepts and images. In this approach, the correspondence problem is attacked as the translation of visual features into concepts, analogous to the statistical machine translation.

5.2.1 Motivation

In the image and video collections, the images are usually annotated with a few keywords which describe the images. However, the correspondences between image regions and words are unknown. For example, for an image showing a zebra on the grass, and having the annotated keywords `zebra` and `grass`, it is known that zebra and grass are in the image, but it is not known which region is zebra and which region is grass (Figure 4). With a single image, it is not possible to solve the correspondence problem. However, if there were other images, where the black and white stripey region (the region corresponding to zebra) was not associated with a green region (which correspond to grass) but with something else (e.g. a gray region corresponding to ground, or a blue region corresponding to sky), then it would be possible to learn that `zebra` corresponded to the black and white stripey region but not to the green one.

This correspondence problem is very similar to the correspondence problem faced in the statistical machine translation literature (Figure 5). There are several parallel corpora (sometimes known as `aligned bitext`), which consist of many small blocks of text in two languages, that are known to correspond to each other at the paragraph or sentence level, but word to word correspondences are unknown.

Brown *et.al* [2] suggested that it may be possible to construct automatic machine translation systems by learning from such large datasets. Using these aligned bitexts, the problem of lexicon learning is transformed into the problem of finding the correspondences between words of different languages, which can then be tackled by machine learning

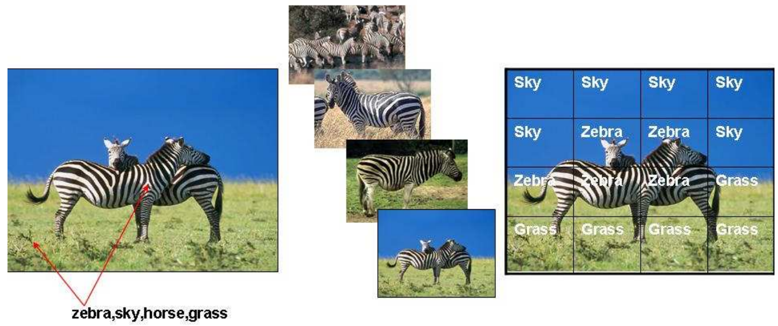

Figure 4: The correspondence problem between image regions and words: The words `zebra`, `grass`, and `sky` are associated with the image, but the word-to-region correspondences are unknown. If there are other images, the correct correspondences can be learned and used to automatically label each region in the image with annotated keywords

methods. In this paper, we explore some ideas from Machine translation for relating words with pictures.

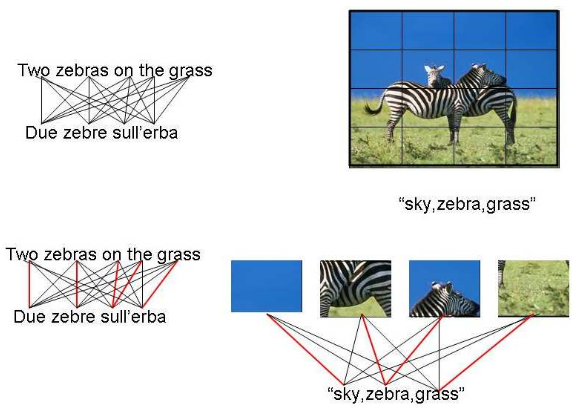

Figure 5: Correspondence problem between image regions and concepts can be attacked as a problem of translating visual features into words. The problem is very similar to Statistical Machine Translation. We want to transform one form of data (image regions or English words) to another form of data (concepts or French words)

Due to this similarity between the problems, the correspondence between image regions and concepts can be attacked as a problem of translating visual features into words, as first proposed by Duygulu *et.al.* [5]. Given a set of training images, it is possible to create a probability table that associates words and visual features which can be then used to find the corresponding words for the given test images.

5.2.2 Approach

In machine translation, a lexicon links a set of discrete objects (words in one language) onto another set of discrete objects (words in the other language). Therefore, in order to exploit the analogy with machine translation, both the images and the annotations need to be broken up into discrete items. The concept annotation keywords in the TRECVID data set can be directly used as discrete items.

In order to obtain the discrete items for visual data, the images are first segmented into regions. The regions could be obtained by a segmentation algorithm as in [5] or can be fixed sized blocks as we will use in this paper. Then, a set of features, such as color, texture, and edge, are computed to represent each region. Finally, the regions are classified into region types (`visterms`) using K-means to perform vector quantization.

After having the discrete items, an aligned bitext, consisting of the visterms and the concepts for each image is obtained. The problem is then, to use the aligned bitext in training to construct a probability table linking visterms with concepts. In this paper, we use the direct translation model. Brown *et. al.* [2] propose a set of models for statistical machine translation. The simplest model (Model 1), assumes that all connections for each French position are equally likely. This model is adapted to translate visterms to concepts, since there is no order relation among the visterms or concepts in the data.

The word posterior probabilities for each visterm, supplied by the probability table, is then used to predict concepts for the test data. In order to obtain the word posterior probabilities for the whole image, the word posterior probabilities of the regions in the image, provided by the probability table, are marginalized as given below:

$$P_0(\mathbf{c}|\mathbf{d_v}) = 1/|\mathbf{d_v}| \sum_{\mathbf{v}\in\mathbf{d_v}} P(\mathbf{c}|\mathbf{v}) \tag{23}$$

where $\mathbf{v}$'s are the visterms in the image. Then, the word posterior probabilities are normalized.

We note that machine translation models that incorporate word order and alignment information did not perform as well as the basic direct translation approach attempted here. This is perhaps due to the lack of any discernable word order in the annotations associated with images.

5.2.3 Comparison with other approaches for relating words with pictures

We note that we implemented the Cross-Media Relevance Models suggested by Manmatha et al. [11, 6, 14]. In particular, we implemented the model with continuous visual features[14]. In addition, we also implemented a Hidden Markov Model for visual concept annotation, using continuous visual features. Figure 6 illustrates the HMM topology used in our experiments and detailed in Ref.[8]. We note that both these models have better visual annotation performance compared to the Machine Translation approach. However, these improvements in annotation performance did not translate to gains in Information Retrieval task (Section 6). The best IR gains that we have been able to achieve so far has been with the use of MT models for visual concept annotation. We are currently investigating the cause of this behavior. Figure 7 presents the comparison between these three approaches for visual concept annotation.

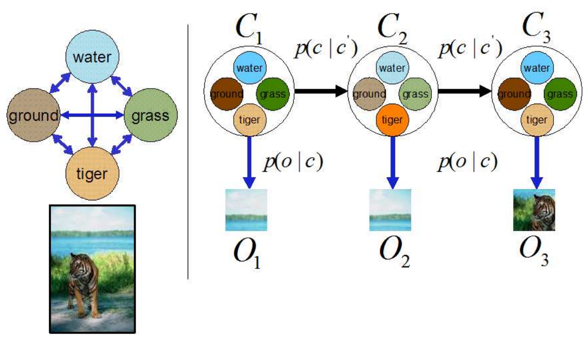

Figure 6: The Hidden Markov Model topology used in visual concept annotation experiments in this paper. The states represent the annotation words and the observations are feature vectors from a grid partitioning of the image, as in the case of Machine Translation models

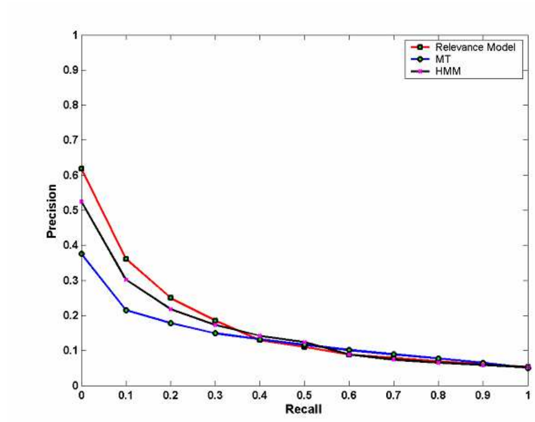

Figure 7: Comparison of the MT models with HMM and Relevance Models for visual concept annotation

5.3 Relating the Query Visual Representation with the Document

In this section we detail the experiments conducted to relate the visual part query to the ASR text of the documents (i.e. modeling $p(\mathbf{c}|\mathbf{d_w})$). As before, we adopt an information bottleneck approach and build models for extracting visual concepts from the ASR text. We use the same set of concepts used in the visual models and therefore these experiments are comparable in that sense, albiet they operate on different parts of a multimedia document.

5.3.1 *Preprocessing the ASR*

Rather than using the raw ASR output available with the TRECVID corpus, we applied the following preprocessing steps to extract textual features from ASR.

- **Sentence boundary creation**. We used a simple approach that recursively segmented the ASR text based on hypothesized silence duration till, on average, the segments had 20 words. This simple approach seemed to work reasonably well. We also used *mxterminator* to break the ASR text into sentences[20]. Our initial assessment indicates that this did not make a significant difference to the final performance of the text processing pipeline. In the final experiments we used the simple scheme outlined above. We note that we get clause boundaries rather than true sentence boundaries but this does not seem to affect later processing.

- **Case restoration**. The segmented ASR was then passed through a case restorer to recover as much of the case information as possible. The case restorer has a built-in lookup table for proper names.

- **Part-of-Speech Tagging and Named Entity Extraction**. The case-restored text was input in parallel to *mxpost*[19] for Part-of-Speech tagging and to *annie*[3] for Named Entity tagging. Only the nouns are retained from the POS tagging.

- **Wordnet processing**. The extracted nouns are filtered using Wordnet and any abstract nouns are removed[7]. The remaining nouns are expanded with their hypernyms and the expanded list is filtered to allow only those nouns that are in the concept list.

At the end of this process, we obtain about 7000 unique word tokens. This list is expanded by considering tokens from the previous and next shot around a shot of interest. This expansion was based on the assumption that there is typically a mismatch between the spoken text and the visuals. In addition, our experiments (description follows) indicated that expanding to one neighbor on each side was sufficient. This increases the total words to about 18000 unique tokens. All further experiments are based on this token set.

5.3.2 *Naive Bayes, SVM and MaxEnt classifiers for Concept Annotation*

We built Naive Bayes and SVM classifiers using this standard token set. The classifiers were binary classifiers, one for each concept (presence/absence). We also performed per classifier feature selection. For each classifier, we use Mutual Information to select the number of word tokens and then train each classifier. This results in a significant reduction in the number of text features required for each classifier model. The MaxEnt classifier differed from the SVM and Naive Bayes classifiers in that it was a multi-way classifier. Likewise, one set of optimal features for all the 75 concepts were chosen using Mutual Information for the MaxEnt model. For these classifiers, the text features were binary, that is if a word token is present, the feature is considered present. We used the Weka machine learning system for implementing the Naive Bayes and SVM classifiers and the openNLP toolkit for MaxEnt modeling[23, 10].

5.3.3 The Language Model Based Classifier for Concept Annotation

The language model (LM) based classifier trains two language models on the training data. One on the set where the concept is present and the other one on the part of the data where the concept is absent. During testing, both language models are used to calculate perplexity on the test data. The one which gives the smaller perplexity determines the concept assigned to the test data. In principle this is a variant of a Bayes classifier, using feature counts as opposed to binary presence/absence values. Formally that corresponds to

$$\text{argmax}_{\mathbf{c}\in\mathbf{c}_{present},\mathbf{c}_{absent}} \prod_i P(f_i|\mathbf{c})P(\mathbf{c})^{\gamma} \tag{24}$$

where f_i are the feature from the test set and γ corresponds to the "Language Model Factor" in speech recognition. The probabilities of the language models are smoothed using absolute discounting:

$$P(f_i|\mathbf{c}) = \max(\frac{N(f_i,\mathbf{c})-d}{N(\mathbf{c})},0) + \frac{dR}{N(\mathbf{c})} \tag{25}$$

with $R = \sum_{i:N(f_i,c)>0} 1$ and d the discounting parameter. Note that $P(\mathbf{c})$ does not need any smoothing. The two parameters γ and d are optimized on a validation set.

Fig. 8 shows the relative frequency of the 75 concepts used in this paper. The concepts are sorted by their frequency. Note that only the y-axis is logarithmic and that the data is best fitted by an exponential. Naively, one might think, that the number of concepts determines the information that can be passed from the visual to the textual models. However, a closer inspection of the plot shows, that only a few concepts contribute significantly. The self perplexity is 29.7. This is still a relatively high value. Given the fact that each shot has on average 3.8 annotations (from the list of 75 concepts) the set of annotations can still give a relatively accurate account of the content of the image.

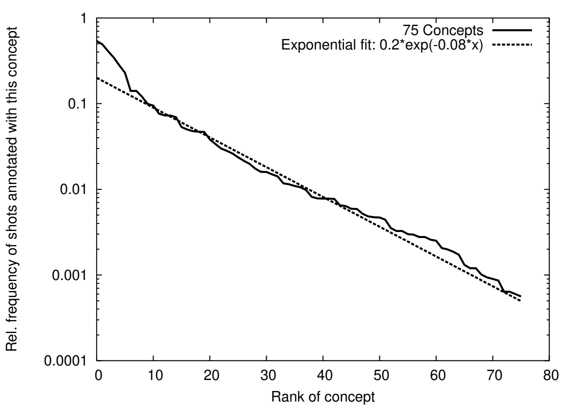

Figure 8: Relative frequency of the 75 concepts sorted by their frequency

Table 1 presents the comparison between the 4 text-based concept annotation approaches investigated in this paper. We note that the difference between the LM approach and the SVM approach is not statistically significant. The difference between the MaxEnt models and the LM models are significant at a p-value of 0.01.

	Chance	LM	SVM	NB	MaxEnt
mAP	0.050	0.125	0.116	0.102	0.100

Table 1: Comparison of the different methods to extract concepts from ASR. mAP corresponds to Mean Average Precision, a figure-of-merit used by NIST to evaluate TREC IR systems. It is the mean of average precisions for all queries used to evaluate a system. Average Precision is the ratio of the sum of precisions at all relevant documents to the total number of relevant documents in the corpus for that query.

6. INFORMATION RETRIEVAL EXPERIMENTS ON THE TRECVID03 CORPUS

In the previous sections we described the different component models that capture the relationship between words and visual terms in a multimedia document. We measured the effectiveness of these models using a set of 75 semantic concepts that was made available as part of the TRECVID03 development corpus. In this section, we integrate the various components models together using the two retrieval models outlined in Section 3, namely the linear and the log-linear models. We employ the TRECVID03 search set and queries for evaluating these retrieval models[16].

The baseline model is a text-only retrieval model which is modeled as $p(\mathbf{q_w}|\mathbf{d_w})$. First, we combine the baseline model with document visuals using the Machine Translation approach outlined in Section 5.2 (and labeled MT in the table 2 below). With this model, the query words are now related both the document words and document visterms. In the linear retrieval case, this corresponds to the following retrieval equation.

$$p(\mathbf{q_w}|\mathbf{d_w},\mathbf{d_v}) \approx \lambda_w p(\mathbf{q_w}|\mathbf{d_w}) + (1-\lambda_w)p(\mathbf{q_w}|\mathbf{d_v}) \tag{26}$$

Similarly, the corresponding log-linear model is

$$p(\mathbf{q_w}|\mathbf{d_w},\mathbf{d_v}) \approx p(\mathbf{q_w}|\mathbf{d_w})^{\lambda_1} \times p(\mathbf{q_w}|\mathbf{d_v})^{\lambda_2} \tag{27}$$

In both cases, the component $p(\mathbf{q_w}|\mathbf{d_v})$ is modeled using the information bottleneck layer (see Section 5) and therefore it expands to the following equation:

$$p(\mathbf{q_w}|\mathbf{d_v}) = \sum_{\mathbf{c}} p(\mathbf{q_w}|\mathbf{c})p(\mathbf{c}|\mathbf{d_v}) \tag{28}$$

In these experiments, we model $p(\mathbf{q_w}|\mathbf{c})$ and $p(\mathbf{q_v}|\mathbf{c})$ with a simplified query-concept model, namely $p_q(\mathbf{q}|\mathbf{c})$ which is derived adapting $p(\mathbf{q_w}|\mathbf{c})$ on a set of development queries outside the TRECVID03 collection.

When we combine the Concepts from ASR model into the retrieval framework, we get a combined equation for the log-linear model as shown below. The linear model is derived similarly.

$$p(\mathbf{q_w}|\mathbf{d_w},\mathbf{d_v}) \approx p(\mathbf{q_w}|\mathbf{d_w})^{\lambda_1} \times p(\mathbf{q_w}|\mathbf{d_v})^{\lambda_2} \times p(\mathbf{q_v}|\mathbf{d_w})^{\lambda_3} \tag{29}$$

In Fig. 9 we compare linear with log-linear interpolation as a method to combine the different models. It is slightly inappropriate to compare the linear fusion model with the log-linear model in the same graph since the linear model weights are limited to be in the range $0-1$ and the log-linear models do not share the same restriction and the interpolation weights have a different meaning for these models.

However, the main emphasis here is to illustrate that the log-linear model outperforms the linear model and that there are no suitable weights for which the linear model improves over the baseline model. Further, there are regimes of interpolation weights where the log-linear model significantly outperforms the baseline. In the graph, a combination of the baseline model with the machine translation model for concept annotation is shown. The difference between the two methods is surprisingly large. In language modeling for speech recognition, we observe that log-linear interpolation is better than linear interpolation but the difference is never as large as in this figure. Our working hypothesis is that the difference between the information retrieval performance of the text modality and the visual modality is very large[2]; The log-linear model is better able to handle this disparity in performance of the respective modalities. Given that the log-linear model is derived using Maximum Entropy principles, it aims to have a smaller bias in the resulting probability density function of the joint model. This smaller bias in the pdf is perhaps responsible for the superior performance of the log-linear model compared with the linear model, especially when one of the modalities overwhelms the others in performance.

In addition, it is striking, that there is no interpolation weight where linear interpolation gives a benefit. This may be due to the fact that we could have a problem in converting our concept annotation models into proper retrieval probabilities with a reasonable distribution of the probabilities in the interval [0 : 1].

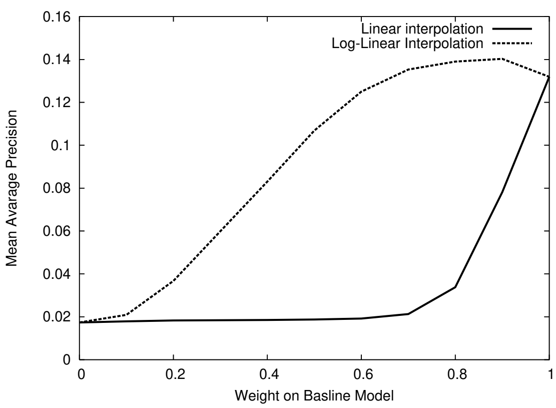

Figure 9: Comparison of the two methods to combine models

In Table 2 we give the results of the fusion experiments. First, we add the MT system outlined in Section 5.2. Then, we add in the ASR-based concept system (Section 5.3).

Finally Fig. 10 gives the recall-precision curve of the overall best model, a combination of the baseline with the machine translation model for image annotation and the model that extracts concepts from ASR. We note that these results are signficant at the 95% level using a paired t-test. We observe that we get a consistent improvement in the high-precision region.

[2]as has been noted by several other researchers, the visual modality is atleast an order of magnitude lower in IR performance compared to the text modality on the same corpus cf. [16]

Model	Retrieval mAP
Baseline	0.131
+ MT	0.139
+ Concepts from ASR	0.149

Table 2: Results from fusing the different joint visual-text models. Note that the final fusion result is significantly superior to the Baseline (at 95% level) using paired t-test.

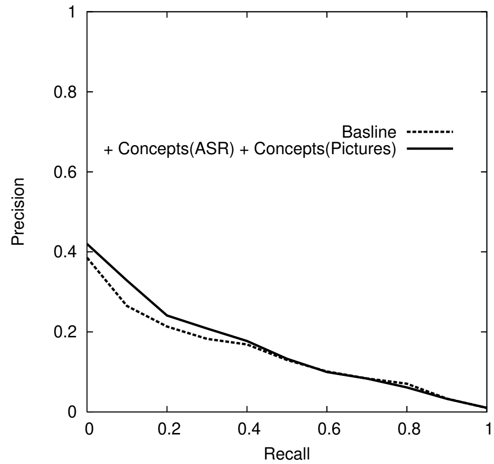

Figure 10: Recall-Precision comparison between the baseline and the final fusion system

7. SUMMARY AND FUTURE WORK

In this paper, we investigated a novel approach for multimedia retrieval which jointly models the visual and textual components of a video shot. In particular, we presented a retrieval framework where individual components for modeling the different aspects of the query and document interaction can be plugged into an overall system. We built automatic multimedia retrieval systems using this approach. We proposed two fusion models – linear and log-linear. The log-linear model is inspired from a Maximum Entropy formulation and our experiments indicate that this model has a superior fusion performance on our experimental corpus. Experiments were conducted on the TRECVID03 corpus and initial results indicate that we get a 14% improvement in retrieval performance using joint models over a text-only baseline. We illustrated the novel framework by building several components that relate different parts of the query with different parts of the multimedia document. The framework is flexible and permits several other techniques that relate different parts of multimedia documents and queries to be combined into a unified whole.

In particular, we used a Machine Translation inspired approach for relating the visual part of the document to the text part of the query. This approach extends the work done by Duygulu et al[5] to the TRECVID03 corpus. We observed that the direct translation approach works best for concept detection and annotation. To relate the visual part of the query to the ASR text of the video shot, we investigated several approaches for extracting visual concepts from ASR text, including MaxEnt models, Naive Bayes models and unigram count based models. These approaches indicate that predicting visual concepts from ASR, while a challenging and counter-intuitive task, does appear possible and perhaps even competitive to visual-only approaches. However, it is not clear what is the upper-limit on performance of such an approach.

Some of the challenges that we faced with this corpus included incomplete labeling of images (i.e. only a few concepts were marked in the images and not all the ones that were present). Also, these annotations were conducted by a large group of people (see NIST TRECVID common annotation forum[16]) and the quality varied significantly between annotators. We did not exploit any spatial or temporal dependencies in our experiments. This needs to be better explored in future work. Also, expanding the size of the bottleneck and perhaps direct modeling of queries and documents needs to be explored. In our experiments, very little query dependent processing was attempted. We note from literature that such techniques have worked well for several IR tasks. This is an important future direction for further performance improvements. One of the streams of information that we did not exploit in these experiments include on-screen text. Our assessment indicated that this information is very relevant for many queries. However, off-the-shelf optical character recognition (OCR) programs perform poorly on such images and produce significantly degraded text. If the quality of video OCR output can be improved, this source of information will become quite useful and can be easily integrated into the approaches that we developed here.

8. ACKNOWLEDGEMENT

This is material is based upon work supported by the National Science Foundation under Grant No. 0121285. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

[1] A. Berger and J. Lafferty. The Weaver System for Document Retrieval. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pages 163–174. NIST Special Publication 500-246, 2000.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Linguistics*, 19(2):263–311, 1993.

[3] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 168–175, 2002.

[4] J. Darroch and D. Ratcliff. Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[5] P. Duygulu, K. Barnard, N. de Frietas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Lecture Notes in Computer Science*, 2353:97, 2002.

[6] S. L. Feng, , R. Manmatha, and V.Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Intl. Conf. on Computer Vision and Pattern Recognition*, Washington D.C., June 2004.

[7] G. A. Miller. WordNet: A Lexical Database. *Communications of the ACM*, 33(11):39–41, 1995.

[8] A. Ghoshal, P. Ircing, and S. Khudanpur. Hidden Markov Models for Automatic Image Annotation and Content-based Retrieval of Images. In *Proceedings of the Twenty-Eigth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Brazil, 2005. ACM Press.

[9] A. Hauptmann, D. Ng, R. Baron, M. Chen, and et. al. Informedia at TRECVID 2003: Analyzing and searching broadcast news video. In *Proceedings of TRECVID2003*, Gaithersburg, MD, November 2003. NIST.

[10] T. M. J. Baldridge and G. Bierner. openNLP maximum entropy modeling toolkit. http://maxent.sourceforge.net/, version 2.2.0, 2004.

[11] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the Twenty-Sixth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 119–126, Toronto, Canada, 2003. ACM Press.

[12] D. Klakow. Log-linear interpolation of language models. In *Proc. International Conference on Speech and Language Processing (ICSLP*, Sydney, Australia, November 1998.

[13] J. Lafferty and C. Zhang. Document language models, query models, and rish minimization for information retrieval. In *Proceedings of the Twenty-Fourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119, New Orleans, Louisiana, 2001.

[14] V. Lavrenko, S.L.Feng, and R. Manmatha. Statistical models for automatic video annotation and retrieval. In *Intl. Conf. On Acoust., Sp., and Sig. Proc.*, pages 417–420, Montreal, QC, May 2004.

[15] C.-Y. Lin, B. Tseng, and J. R. Smith. Video Collaborative Annotation Forum: Establishing Ground-truth Labels on Large Multimedia Datasets. In *Proceedings of the TRECVID2003: NIST Special Publications*, Gaithersburg, MD, 2003. NIST.

[16] NIST. *Proceedings of the TREC Video Retrieval Evaluation Conference(TRECVID2003)*, Gaithersburg, MD, November 2003.

[17] NIST. *Proceedings of the TREC Video Retrieval Evaluation Conference(TRECVID2004)*, Gaithersburg, MD, November 2004.

[18] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the Twenty-First Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia, 1998. ACM Press.

[19] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In E. Brill and K. Church, editors, *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 133–142. Assn Comp. Ling., Somerset, New Jersey, 1996.

[20] J. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington DC, 1997.

[21] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[22] T. Westerveld and A. P. de Vries. Multimedia retrieval using multiple examples. In *Proceedings of Conference on Image and Video Retrieval CIVR*, Dublin, Ireland, July 2004.

[23] I. H. Witten and E. Frank. *Data Mining: Practical machine learing tools with Java implementations*. Morgan Kaufmann, San Mateo, CA, 1999.

[24] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Neo, and T.-S. Chua. VideoQA: Question answering on news video. In *Proceedings of the ACM Multimedia Conference*, Berkeley, CA, November 2003. ACM.