

# Exploiting Redundancy in Cross-Channel Video Retrieval

Bouke Huurnink  
ISLA, University of Amsterdam  
Kruislaan 403  
Amsterdam, The Netherlands  
bhuurnin@science.uva.nl

Maarten de Rijke  
ISLA, University of Amsterdam  
Kruislaan 403  
Amsterdam, The Netherlands  
mdr@science.uva.nl

## ABSTRACT

Video producers, in telling a news story, tend to repeat important visual and speech material multiple times in adjacent shots, thus creating a certain level of redundancy. We describe this phenomenon, and use it to develop a framework to incorporate redundancy for cross-channel retrieval of visual items using speech. Testing our models in a series of retrieval experiments, we find that incorporating the fact that information occurs redundantly into cross-channel retrieval leads to significant improvements in retrieval performance.

## Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [Information Systems Applications]: H.4.2 Types of Systems; H.4.m Miscellaneous

## General Terms

Algorithms, Measurement, Performance, Experimentation

## Keywords

Speech-based video retrieval, redundancy, language modeling, document expansion

## 1. INTRODUCTION

Millions of people view digital multimedia every day. Precise numbers are volatile and difficult to estimate, but popular video portals YouTube and Dailymotion recently reported 100 million<sup>1</sup> and 16 million<sup>2</sup> respective daily page views. This explosion of video popularity is coupled with an

<sup>1</sup>[http://www.youtube.com/press\\_room\\_entry?entry=jwIToyFs2Lc](http://www.youtube.com/press_room_entry?entry=jwIToyFs2Lc)

<sup>2</sup><http://www.timesonline.co.uk/tol/news/world/article638866.ece>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR'07, September 28–29, 2007, Augsburg, Bavaria, Germany.  
Copyright 2007 ACM 978-1-59593-778-0/07/0009 ...\$5.00.

explosion in the amount of online digital content, provided by large user communities but also by news broadcasters, educational institutions, public archives and museums, to name just a few examples. Commercial navigation tools traditionally rely on metadata such as tags or proximate text to locate the right video for a particular search, an approach which is limited by the extent to which humans are willing to manually annotate each video. This limitation has sparked increased research efforts into *content-based video retrieval*. In the content-based approach, videos are returned based on the images and sound formed by their pixels and audio bits, rather than metadata that happens to be associated with them [12].

Video conveys information through multiple channels, such as speech, music, movement, and images. Each of these channels is temporally governed. As we watch a video segment, we gain understanding of its content by integrating different forms of information over time. News video producers, in order to make the information in video easier to absorb, often build in *redundancy*: the phenomenon that information that is important to the video is repeated several times in multiple shots, both within the same channel and across multiple channels. For example, important subjects on the screen may also be mentioned in the dialogue. In this way, if we are temporarily distracted from listening to the dialogue or looking at the screen, we still have an idea of the semantic content of the video.

Our working assumption is that redundancy of the type described above can be used to improve the effectiveness of video retrieval. In this paper we are especially interested in using redundancy—both within a single channel and across channels—to improve the effectiveness of speech-based retrieval of visual items. To be able to address this issue we first examine the redundancy phenomenon itself. Let us explain. We call a shot *visually relevant* to an object, person, or scene when that particular item can be visually observed in the shot. Now, assuming that we know that a given shot is visually relevant, how likely is it that a neighboring shot is visually relevant as well? And if visual relevance does spread out to neighboring shots, at which distance can you still observe the effect? The phenomenon is illustrated in Figure 1. Here we see keyframes extracted from four consecutive shots in a news broadcast featuring Tony Blair, the former prime minister of the United Kingdom. The keyframes contain similar visual and semantic subject matter, with Tony Blair appearing in both shots 2 and 4. How systematic is this phenomenon?

Let us look at another example of redundancy, this time

time	visual channel	speech channel	visual match?	speech match?
1		he also said that have spread frank with the iraqi issue, he and blair has different chirac said he is wrong to pay tax at the final will focus on the prime minister tonv blair	✗	✓
2		co-operation between the two sides in accordance with the joint communique signed between the two countries will next week egypt to participate in international conference on iraq hopes that the two countries' <small>Lebanese World that is in</small>	✓	✗
3		ararat after passing the palestinian-israeli situation blair applying expressed the hope that palestine and israel to seize the opportunity reopen peace process music of the palestinian <small>elaborate, can be, elaborate hu</small>	✗	✓
4		practice and parts are planned in the middle east establishment of an independent palestinian state and the middle east visit to britain's agenda also include lun time, in the evening 18 british queen windsor castle <small>in the photo reviewed</small>	✓	✗

**Figure 1: Finding Tony Blair: an example of temporal item distribution across the speech and video channels**

across channels. When objects are difficult to detect visually, a retrieval engine can look for clues in other channels. For example, individual people such as Tony Blair or Saddam Hussein tend to be very difficult for a system to detect using only visual information. However, a news broadcast showing Tony Blair is likely to mention his name several times, and many state-of-the-art visual retrieval systems report use of speech in their retrieval and/or concept detection algorithms [3, 2, 15, 4, 1]. An interesting challenge emerges here: the *temporal mismatch* that can occur between channels. While each channel is temporally cohesive in its own right, the content may not be synchronized between them. For example, in Figure 1 we see a displacement between the mention of Tony Blair’s name in shot 1, and his appearance in shot 2. It has been shown that named people are on average mentioned two seconds before they appear in broadcast news [19]. We perform an investigation into the distribution of visual to visual relevance (i.e., how likely it is for visual items to occur close together) and contrast this with the distribution of cross-channel speech to visual relevance (i.e., how likely it is for visual items to occur, given temporal proximity to their mention in the speech channel).

Now, assuming that we have developed some understanding of the redundancy phenomenon (i.e., a distribution to model it), the next step is to try and exploit the phenomenon to improve the effectiveness of speech-based retrieval of visual items. In a cross-channel retrieval system, we could utilise redundancy, and more specifically, the tendency of relevant subject matter to occur close together, by returning temporally related shots at retrieval time. Returning to Figure 1, if speech indicates that shot 4 is visually relevant, there is an increased probability that surrounding shots are also visually relevant. In this paper we find consistent redundancy patterns within and across channels, and we propose retrieval models that integrate these patterns to improve cross-channel retrieval performance.

To sum up, our paper centers on the task of cross-channel video retrieval, using speech to retrieve visual subject matter and the potential of redundancy to improve retrieval effectiveness. This gives rise to the central questions that we focus on in this paper:

1. Given a query, how are visually relevant items distributed in video, and how can we characterize this temporal distribution?
2. How can we characterize cross-channel redundancy while taking into account the frequent temporal mismatch between item occurrences in the speech and visual channels of video?
3. Can we model and use the resulting information so as to improve cross-channel retrieval of visual items using speech?

We address the first two questions by means of a large-scale empirical exploration of real-world video data. We incorporate the results of the exploration in a cross-channel retrieval framework, which we use to test our assumptions about the way in which temporal qualities of video might be used to affect retrieval performance. Our main finding is that we can achieve significant performance gains using redundancy patterns detected in the empirical data distributions.

The remainder of the paper is organised as follows. Section 2 outlines previous work on the temporal distribution of visually relevant material and the temporal mismatch in cross-channel retrieval. Section 3 is exploratory in nature; in it, we describe the redundancy phenomenon.

Section 4 outlines the retrieval framework in which we incorporate various redundancy models, and in Section 5 we describe the retrieval results and analysis. Conclusions are presented in Section 6.

## 2. RELATED WORK

Work on redundancy in video retrieval that is related to this paper can be divided into work investigating visual redundancy and work investigating cross-channel redundancy. Visual redundancy for video has been investigated by Yang and Hauptmann [18], who calculate the transitional probability of a shot being visually relevant given that the previous shot is visually relevant. The visual repetition of the same item in multiple shots has been used in concept detection by Van Gemert et al. [17]. It is worth pointing out that Yang and Hauptmann [18] explore visual redundancy, but only for one adjacent shot; we explore visual redundancy over many consecutive shots.

Some of the most detailed work exploring cross-channel redundancy and the temporal mismatch is that outlined by Yang et al. [19]. They performed a detailed investigation into the occurrence of the names of twenty people in speech, and their appearance in the visual channel of broadcast news. They found that on average a person appears about two seconds after their name is mentioned in the speech of the video. The distribution of each named entity was approximated by a Gaussian model, and the authors found that distributions of people occurring frequently in the news shared similar mean values and standard deviations. They went on to use the distribution models to propagate shot relevance scores, and found that by propagating scores to

adjacent shots according to a time-based distribution they were able to increase MAP by more than 30% over a flat window-based approach. The main difference between our work and the work by Yang et al. [19] is that they only explore cross-channel redundancy for 20 people. As we will see below, we explore cross-channel redundancy for 363 concepts and 94 topics. Furthermore, we use a power law rather than a gaussian to model the distribution, which fits the data very well when considering distribution across shots.

Finally, for our retrieval models we work in the setting of generative language modeling [5, 8, 20]; specifically, to incorporate redundancy models within a language modeling setting, we make use of the document expansion model proposed by Tao et al. [16].

### 3. DISCOVERING REDUNDANCY

Here we describe our exploration addressing the first two of our central questions, (1) *given a query, how can we characterize the temporal distribution of visually relevant items in video?*, and (2) *how can we characterize cross-channel redundancy while taking into account the frequent temporal mismatch between item occurrences in the speech and visual channels of video?*

We start by describing our data set. After that we outline the types of video items that we consider and our methodology for obtaining empirical distributions, before uncovering empirical distributions of redundancy within in the visual channel, and across the visual and speech channel.

#### 3.1 Data Set

The TRECVID [11] benchmark provides us with extensive data about item occurrence within video. We utilize the TRECVID data sets from the 2003–2006 benchmarking evaluations. These test collections yield over 300 hours of English, Arabic, and Chinese news broadcast video. The video from the collections is associated with automatically generated boundary annotations for over 190,000 shots. Also included are ASR transcripts for the datasets, and in the case of Arabic and Chinese videos, machine translations of those transcripts to English. In our experiments, we only consider the (machine translated) English-language transcripts. We further process the transcripts by removing commonly occurring stopwords, and by reducing them to their morphological roots using the Porter [9] stemming algorithm.

A total of 96 official TRECVID topics have been created for the 2003–2006 test sets. Each topic consists of one or two natural language sentences describing the visual content that is desired from relevant shots, as well as multimodal examples. It is also accompanied by a ground truth of relevant shots from the associated test set. After an examination of the data we eliminated two topics, leaving us with a total of 94 topics for evaluation.<sup>3</sup>

Also, recent large-scale efforts have resulted in extensive manual annotation of high-level features, or concepts. We utilize the annotations for concepts made publicly available as part of the MediaMill Challenge [13], as well as those made available by the LSCOM effort [7]. We make use of links between concepts and the WordNet thesaurus created

<sup>3</sup>Topics 0118 and 0119 were eliminated as the (C-SPAN) videos containing the relevant shots were not accompanied by ASR transcripts in that year, rendering it impossible to find the relevant shots using transcripts alone.

by Snoek et al. [14], using the thesaurus synonyms (*synsets*) to create a textual description of each concept. In addition, for those named people that do not have an immediate entry in the thesaurus, we add the name of the person to the concept description. We use the annotation of [14] to remove concepts duplicated between LSCOM and MediaMill, as well as concepts with very few annotations. This results in a combined set of 363 annotated concepts.

#### 3.2 Video Items

As outlined in the previous section, the TRECVID datasets are associated with visual annotations for two different types of items: *topics* and *concepts*. It is important to distinguish between the two, as they have different qualities. Topics are formulated to reflect a searcher’s final information need. Concepts, on the other hand, are formulated to provide a searcher with building blocks to be used in the process of fulfilling his or her information need. As a result, topics tend to be more complicated than concepts. For example, one topic statement is *Find shots with one or more soldiers, police, or guards escorting a prisoner*, while the 363 concepts include *soldier*, *police officer*, and *prisoner*.

Another difference between topics and concepts is in the way they are textually described. The topic statements we use are natural language statements, while we use thesaurus terms for the concepts. Thus, topic statements tend to contain more adjectives and verbs than concept descriptions.

#### 3.3 Methodology

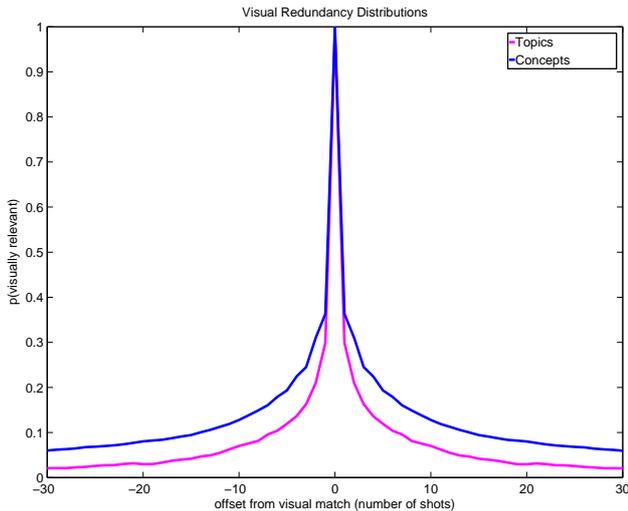
In order to answer our first research question (about the temporal distribution of visually relevant items in video), we have to characterize the redundancy of a visually relevant video item across time. Our approach in answering this question follows the quantitative approach taken by Yang and Hauptmann [18], who are interested in the *transitional probability* of a shot  $e$  being visually relevant to an item, given that the previous shot  $d$  is visually relevant. We extend the approach to include shots more than one step away from  $d$ , in order to allow us to calculate the distribution of transitional probabilities over a shot neighbourhood. The transitional probability is then indicated by  $p(e_n = V_R | d = V_R)$ , where  $V_R$  indicates that a shot is visually relevant, and  $n$  is the number of shots between  $e$  and  $d$ . In cases where  $e$  occurs before  $d$ ,  $n$  is negative. We use the ground truth assessments of  $V_R$  to calculate the transitional probability of an item at offset  $n$  according to

$$p(e_n = V_R | d = V_R) = \frac{c(e_n = V_R, d = V_R)}{c(d = V_R)} \quad (1)$$

where  $c(e_n = V_R, d = V_R)$  is the count of the shot pairs where both  $e_n$  and  $d$  are visually relevant to the item, and  $c(d = V_R)$  is the total number of visually relevant shots in the collection. When  $e_n$  is outside the beginning or end of the video, we smooth with the overall background probability of the item occurring in the test collection.

To answer our second research question (about the temporal mismatch between the speech and visual channels) we calculate the transitional probability of  $e_n$  given that  $d$  has a match in the *speech* channel. Substituting the speech relevance,  $S_R$  of  $d$  into Eq. 1 this gives

$$p(e_n = V_R | d = S_R) = \frac{c(e_n = V_R, d = S_R)}{c(d = S_R)}, \quad (2)$$



**Figure 2: Average probability of a shot being visually relevant, based on the offset from a known visually relevant shot**

where we say that  $d = S_R$  when the speech associated with the shot matches one of the words of the item description. Speech relevance may be assessed in other ways, but in initial experimentation not outlined here we found the binary match method to achieve comparable results to more complicated methods, for example by using cosine similarity.

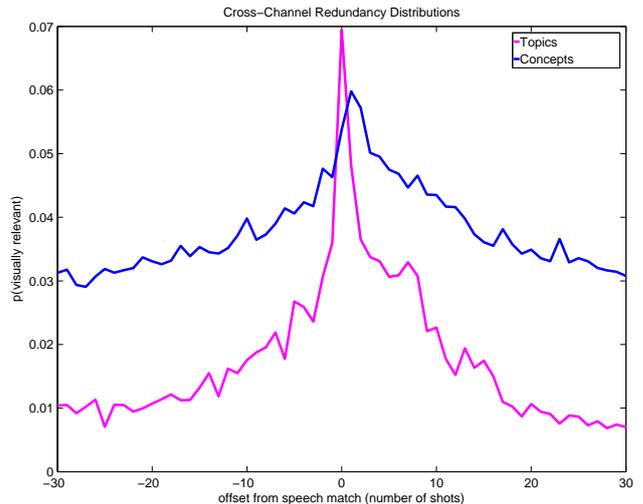
### 3.4 Redundancy in the Visual Channel

Figure 2 addresses the following question: given that the current shot contains a visually relevant item, what is the probability that an adjacent shot is also visually relevant? The graphs are centered around the known visually relevant shot in the middle; along the X-axis we plot the distance from this shot as measured in terms of the number of shots. In Figure 2 concepts and topics are plotted separately, and we see that their curves exhibit a very similar shape. They are both symmetrical, each graph peaks sharply at the shot offset of 0 (the known visually relevant shot), and each graph smooths out to the background probability that any random shot is visually relevant. We see that the concept curve smooths out to a higher background probability than the topic curve. This illustrates that concepts, on average, tend to have more relevant shots in the collection than topics.

### 3.5 Redundancy Across Channels

Figure 3 shows the average transitional probability of visual relevance, given that speech of a shot contains an item word. Noting the scale difference between this figure and Figure 2, we see evidence of cross-channel redundancy for both topics and concepts: there is a clear peak in probability of visual relevance close to the point where a speech match occurs.

Furthermore, we see evidence of a cross-channel mismatch for concepts, where the average transitional probability peaks at the shot after a speech match occurs. Topics do not share this displacement. This is somewhat surprising: topics and concepts are similar in that they both request visual information, and so we might expect them to share any properties of displacement. We speculate that the displacement might



**Figure 3: Average probability of a shot being visually relevant, based on the offset from a known textually relevant shot**

be affected by the more elaborate use of natural language statements for topics, as opposed to (compact) thesaurus synonyms for concepts.

We have now uncovered redundancy phenomena in video retrieval; in the next section we will model these empirical phenomena using power law functions and integrate them in a retrieval framework.

## 4. RETRIEVAL FRAMEWORK

Now that we have uncovered redundancy phenomena in video retrieval, both within a single channel and across channels, we turn to their use for improving retrieval effectiveness. In this section we describe our retrieval framework, and in the next section we put it to use. We start by outlining the language modeling setting in which we work; we follow with a description of a document expansion technique which we will use to incorporate redundancy. To this end we propose models based on power law functions to capture the empirical distributions described in Section 3. We end with a description of the actual inclusion of those models in our retrieval setting.

### 4.1 Retrieval Based on Language Modeling

We base our retrieval framework within the language modeling paradigm. We choose language modeling as it is a theoretically transparent retrieval approach and has been shown to be competitive in terms of retrieval effectiveness [5, 8, 20]. Furthermore, the philosophy behind language modeling fits well with our retrieval wishes. Let us explain.

Our exploration of redundancy has shown that the visual content of a shot is to some extent reflected in the speech of surrounding shots. Therefore, we wish to adjust the speech of each shot with speech from the surrounding shot neighbourhood.

Now, the unigram language modeling approach assumes that a document  $d$  is generated by a random sample of words from a hidden document model  $\theta_d$ , where  $\theta_d$  is a document-specific probability distribution. At retrieval time, for a query  $q$ , each document is ranked with respect to the prob-

ability that  $q$  was generated by  $\theta_d$ . Therefore, the essential problem is estimating  $\theta_d$ . A document is not necessarily a complete reflection of its underlying model, and we can use external information to help estimate  $\theta_d$ . In our approach, we will use speech from surrounding shots to help estimate  $\theta_d$  for each shot.

To give a brief description of the language modeling approach to retrieval, for a query  $q$  containing words  $w_1, w_2, \dots, w_m$ , and a document  $d$  with an estimated model  $\hat{\theta}_d$ , we rank  $d$  according to  $p(q|\hat{\theta}_d)$  so that

$$p(q|\hat{\theta}_d) = \sum_{w \in q} p(w|\hat{\theta}_d).$$

A simple approach to determining  $p(w|\hat{\theta}_d)$  is to use maximum likelihood estimation (MLE). MLE is simply the probability of a word in a document, given by  $\frac{c(w,d)}{|d|}$  where  $c(w,d)$  is the count of  $w$  in  $d$  and  $|d|$  is the total number of words in the document. However, the MLE assigns no probability mass to unseen words, and in addition does not take into account background probabilities of words that occur frequently in the overall document collection. Therefore, some type of *smoothing* is commonly used to adjust for (at least) these factors. In our experiments we use the Jelinek-Mercer smoothing method [20], as we have previously found this to be a suited method for speech-based video retrieval [6]. This method interpolates the maximum likelihood with the background collection language model  $\theta_C$ . The Jelinek-Mercer smoothing estimate is given by

$$p(w|\hat{\theta}_d) = \lambda \cdot \frac{c(w,d)}{|d|} + (1 - \lambda) \cdot p(w|\theta_C), \quad (3)$$

where  $\lambda$  is a fixed parameter that controls the interpolation. This is the model within which we work. Redundancy will be integrated within our retrieval model by adjusting the word counts  $c(w,d)$ , as we will now explain.

## 4.2 Document Expansion

Document expansion is a technique originating from spoken document retrieval that allows for incorporation of external evidence in a natural way [10]. In this approach, a document is expanded and re-weighted with related text at indexing time. Traditionally, this approach is used to augment the original document with text from multiple related documents that have been obtained by some form of feedback. In our approach, document ‘relatedness’ will be assigned according to temporal proximity.

The technique outlined by Singhal and Pereira [10] was specific to the vector space model approach and a certain implementation of relevance feedback. Tao et al. [16] propose a more general model for document expansion, outlined below, on which we build. To perform document expansion, we use a corpus  $E$  to determine additional information about every document  $d$  in a corpus  $C$ . At indexing time we use word counts from  $d$  and from documents in  $E$  to create a ‘pseudo-document,’  $d'$ . The word counts in  $d'$ ,  $c(w,d')$ , are adjusted from those in  $d$  according to:

$$c(w,d') = \alpha \cdot c(w,d) + (1 - \alpha) \cdot \sum_{e \in E} (\gamma_d(e) \cdot c(w,e)), \quad (4)$$

where  $\alpha$  is a constant,  $e$  is a document in  $E$ ,  $\gamma$  is our confidence that  $e$  provides information that is useful for  $d$ , and  $c(w,d)$  is the number of occurrences of  $w$  in  $d$ .

Placing this model in the context of temporally related video data, we have the following:

- our original document is a shot  $d$ , and its associated speech transcript;
- $E$  is defined by the neighbouring shots within a window of  $X$  shots;
- $\gamma$  is the expectancy of speech from  $e$  describing visual subject matter contained in  $d$ , given its offset  $n$  from  $d$ ;
- $d$  is by definition the central member of  $E$ , and this eliminates the need for  $\alpha$ , which is replaced by the  $\gamma$  value at offset 0.

This leads to a simplified expanded document model:

$$c(w,d') = \sum_{e \in E} (\gamma_d(e) \cdot c(w,e)), \quad (5)$$

Below, we arrive at different retrieval models by making different choices for  $\gamma$ ; then, Eq. 5 is used instead of the original word count  $c(w,d)$  in Eq. 3.

## 4.3 Modeling Expectancy of Visual Relevance

Now, we need to determine the expected visual relevance  $\gamma$ . We develop empirical models for  $\gamma$  directly from the empirical distributions found at the end of Section 3. In addition, we use power law functions to approximate the empirical distributions in an attempt to find a simplified model for  $\gamma$ .

Modeling  $\gamma$  from the empirical distributions, we normalize the probabilities of each distribution so that  $\gamma = 1$  at the maximum value, and  $\gamma = 0$  when the transitional probability is equal to the background probability of the item occurring anywhere in the test collection. The results can be seen in the data points of Figures 4, 5, 6 and 7.

Furthermore, we use logistic regression on the data points to develop power law models of the form  $\gamma = bx^m$ , where  $b$  and  $m$  are constant, and  $x$  is the absolute offset from the shot with the highest visual relevance probability. In the case of cross-channel redundancy, where the data-points are asymmetrical on either side of the centre, we regress a separate power law function for each side of the curve. The regressed power law curves are plotted in Figures 4, 5, 6 and 7, and each curve is labelled with the regressed formula values. Interestingly, the curves for the visual redundancy distributions are very similar to the simple natural harmonic series  $\{\frac{1}{1}, \frac{1}{2}, \dots, \frac{1}{x}\}$ , or  $x^{-1}$ . This is especially true for the function describing concept redundancy, with the equation  $0.9515x^{-1.0101}$  approaching that for the harmonic series. In fact, the power law function shown in Figure 5 fits the harmonic series so well that it largely obscures the harmonic plot.

## 4.4 Integrating Redundancy

Finally, we need to put together the two main ingredients developed so far: our retrieval framework and the various choices for the expected visual relevance listed above.

We integrate redundancy into our framework by modifying the  $\gamma$  function in Eq. 5. We consider five variations, yielding five retrieval models (in addition to the baseline), which we now describe:

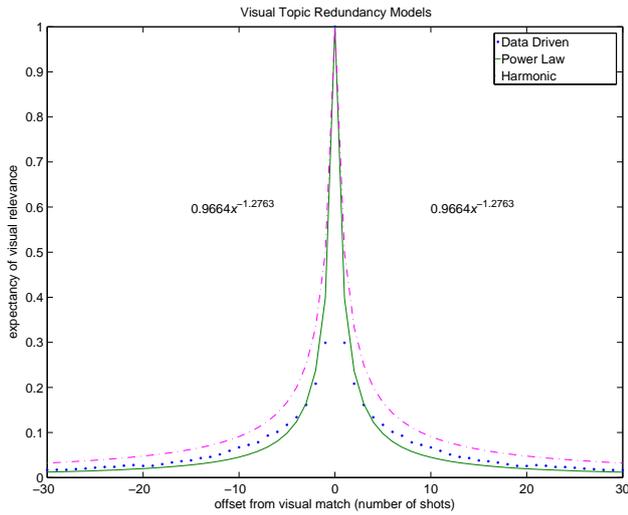


Figure 4: Expected visual relevancy for topics, based on visual redundancy

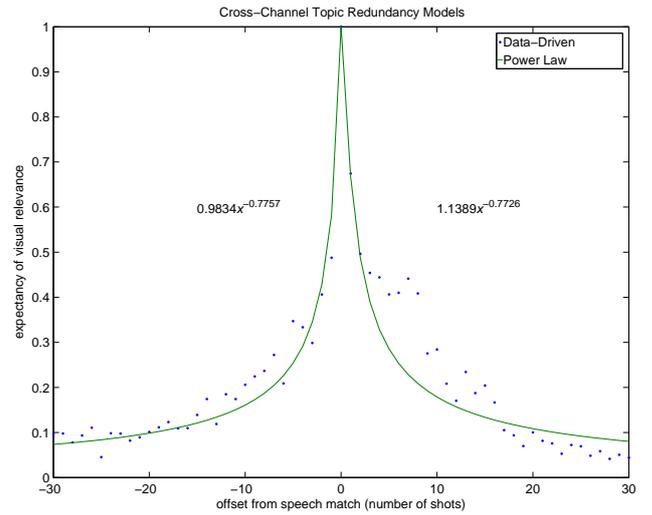


Figure 6: Expected visual relevancy for topics, based on cross-channel redundancy

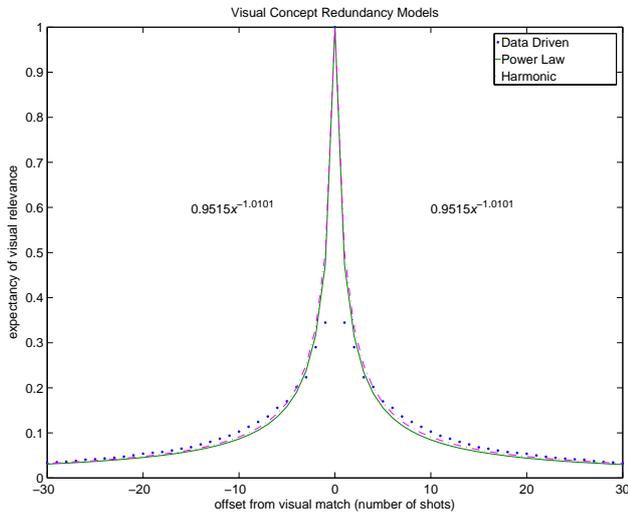


Figure 5: Expected visual relevancy for concepts, based on visual redundancy

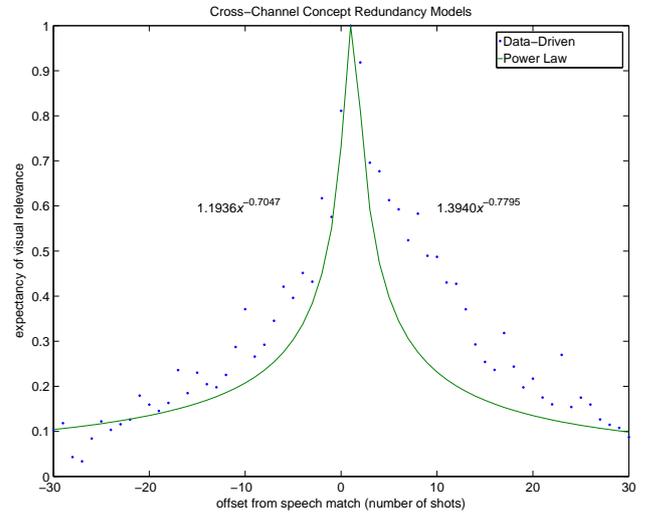


Figure 7: Expected visual relevancy for concepts, based on cross-channel redundancy

0. **baseline** no integration of redundancy; uses the model described in Eq. 3;
1. **flat**  $\gamma = 1$ : all shots are expected to be equally visually relevant;
2. **speech data driven**  $\gamma$  is determined by the empirical cross-channel redundancy value at offset distance  $n$ ;
3. **visual data driven**  $\gamma$  is determined by the empirical visual redundancy value at distance  $n$ ;
4. **speech model driven**  $\gamma$  is determined by a power law approximation of cross-channel redundancy at distance  $n$ ;
5. **visual model driven**  $\gamma$  is determined by a power law approximation of visual redundancy at distance  $n$ .

## 5. RETRIEVAL EXPERIMENTS

In our retrieval experiments we use the retrieval framework developed in Section 4 to address the following questions:

- a) Can visual redundancy patterns be used to improve cross-channel retrieval performance?
- b) Can cross-channel redundancy patterns be used to improve retrieval performance?

We test each of the retrieval models described in Section 4.4 using the data set and items described in Section 3. The models are evaluated at increasing shot window sizes. As our baseline we use a search using shot speech only (i.e., not expanded to include speech from neighbouring shots).

We also have some subsidiary research questions for which we seek answers in this section:

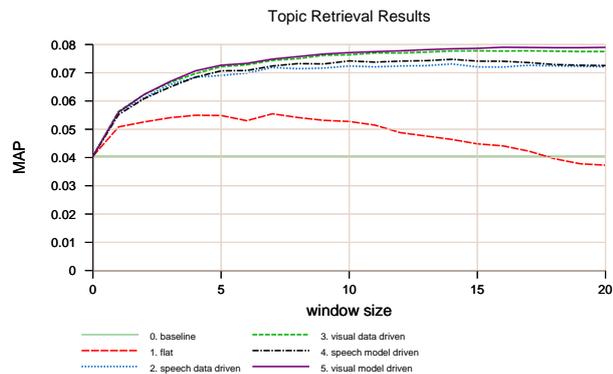


Figure 8: Topic MAP scores for different retrieval models, using different window sizes

- c) Do topics and concepts differ—in performance and, more interestingly, in the effectiveness achieved by the retrieval models that we consider?
- d) Is there a difference in performance between the empirical models on the one hand and the power law based models on the other?

Results are evaluated using the standard Mean Average Precision (MAP) measure. We give an overview of our results and describe them in further detail below.

## 5.1 Result Overview

Figures 8 and 9 provide MAP curves for each of the retrieval models that we consider, for both topics and concepts, respectively. We see that redundancy based methods (models 2, 3, 4 and 5) consistently—i.e., for all window sizes—outperform both the baseline (model 0) and the flat method (model 1). In general we can observe that the MAP scores initially increase as the window size increases, and that they level off at a window size of around 10 for topics, and around 15 for concepts. We also see that the performance of the two cross-channel “speech models” (models 2 and 4) is very similar and that the performance of the two “visual models” (models 3 and 5) is very similar. We see that topics achieve better performance with models based on visual redundancy, while concepts perform better with models based on cross-channel redundancy. Furthermore, we observe that models based on power law approximations perform comparable to those based on empirical data points.

Table 1 provides an overview of the MAP scores achieved by the baseline and the other five retrieval models that we consider. Other than the baseline, the scores in the table are all at a window size of 20 shots to either side of the current shot.

## 5.2 Significance Tests

In Table 2 we show the values for significance test between the scores of the different retrieval models, once again at a window size of 20 shots. Significance testing was done using the Wilcoxon matched-pairs signed-ranks test, at the 0.05 level.

The redundancy based models all performed significantly better than model 0 (baseline) and model 1 (flat). Model 1

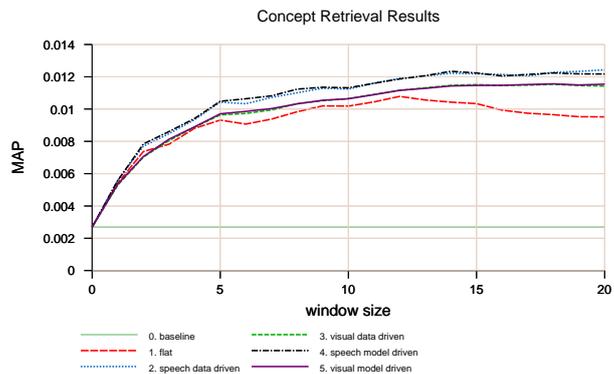


Figure 9: Concept MAP scores for different retrieval models, using different window sizes

Table 1: MAP scores for topics and concepts for the baseline, and for the retrieval models at window size 20. The highest score for each item type is highlighted in bold face.

Retrieval model	Topic MAP	Concept MAP
0. baseline	0.0405	0.0027
1. flat	0.0373	0.0095
2. speech data driven	0.0721	<b>0.0124</b>
3. visual data driven	0.0775	0.0114
4. speech model driven	0.0726	0.0122
5. visual model driven	<b>0.0790</b>	0.0115

in turn performed significantly better than the baseline for concepts, however for topics it did not.

Looking at the differences between the redundancy based models, we see that there are few significant differences for concept retrieval results, likely due to the small magnitude of the MAP scores. For topics, on the other hand, we see the visual redundancy models 3 and 5 significantly outperforming the cross-channel redundancy models 2 and 4.

## 5.3 Differences between Topics and Concepts

While one should be careful comparing MAP scores across different sets of queries, it is interesting to observe that the retrieval performance differs substantially between topics and concepts. Firstly, we find that our retrieval framework produces consistently higher MAP scores for topics than it does for concepts. As shown in Table 2, at window size 20 the highest score for topics is almost 6 times larger than for concepts, with MAP scores of 0.0790 and 0.0124 respectively. This holds across retrieval models, and indicates that cross-channel retrieval generally works better for topics than for concepts. In other words, topic words are more likely to be mentioned in speech than concept words, which stands to reason, as concepts tend to be expressed more concisely than topics (with 4.6 vs 3.5 words on average, after stop word removal). Informal analysis also indicates that topics may be more likely to be mentioned in speech than concepts, with a higher fraction of requests for specific people, places, and things.

Secondly, we find that topics and concepts differ in the types of models that achieve the best performance. Topics

**Table 2: Pairwise comparison of retrieval models that integrate redundancy information. Significant differences are in boldface. (Single digits in columns and rows indicate retrieval models; 0: baseline; 1: flat; 2: speech data driven; 3: visual data driven; 4: speech model driven; 5: visual model driven.)**

Concepts		0	1	2	3	4
1	<b>0.0000</b>					
2	<b>0.0000</b>	<b>0.0000</b>				
3	<b>0.0000</b>	<b>0.0000</b>	0.8671			
4	<b>0.0000</b>	<b>0.0000</b>	0.1658	0.1459		
5	<b>0.0000</b>	<b>0.0000</b>	0.3949	<b>0.0021</b>	0.09614	

Topics		0	1	2	3	4
1	0.3439					
2	<b>0.0001</b>	<b>0.0001</b>				
3	<b>0.0000</b>	<b>0.0000</b>	<b>0.0100</b>			
4	<b>0.0000</b>	<b>0.0000</b>	0.2488	<b>0.0061</b>		
5	<b>0.0000</b>	<b>0.0000</b>	<b>0.0063</b>	0.1038	<b>0.0015</b>	

achieve the highest performance using models based on visual redundancy, while concepts perform consistently better under models based on cross-channel redundancy. We observed in Section 3.5 that the visual occurrence of concepts is displaced to one shot after their mention in speech. The cross-channel redundancy model takes this into account, and the results suggest that this is beneficial for retrieval. Topics, on the other hand, are overall less affected by cross-channel displacement, with visual relevance peaking at the same shot where a topic is mentioned in speech. In this case, the visual redundancy patterns seem to be a better descriptor of visual relevance than cross-channel redundancy.

#### 5.4 Power Law Approximations versus Empirical Models

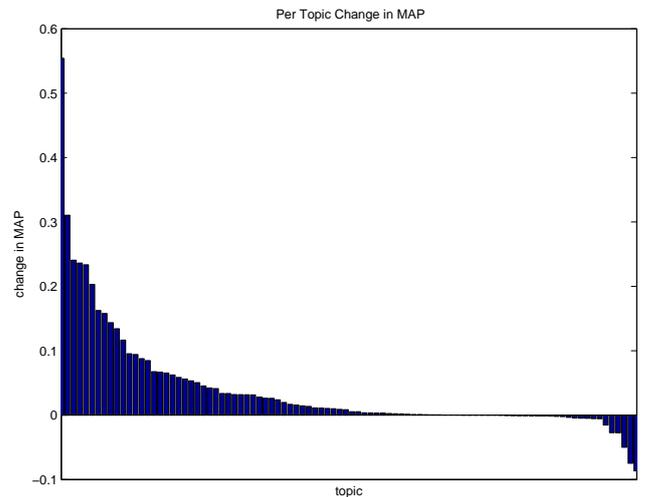
Retrieval models 4 and 5 are based on power law functions that approximate the empirical data points used in models 2 and 3. For topics we find that the power law based retrieval models outperform the ones based on empirical models, although the differences are not significant. For the concepts a mixed message emerges: the empirical model has a slight edge over the power law model in one case (model 2 vs model 4) and vice versa in the other case (model 3 vs model 5).

In sum, the significance tests in Table 2 show that the retrieval models based on power law approximations do not produce significantly lower MAP scores than those based on empirical data.<sup>4</sup> Hence, it seems safe to recommend the use of power law based models for incorporating redundancy within retrieval models.

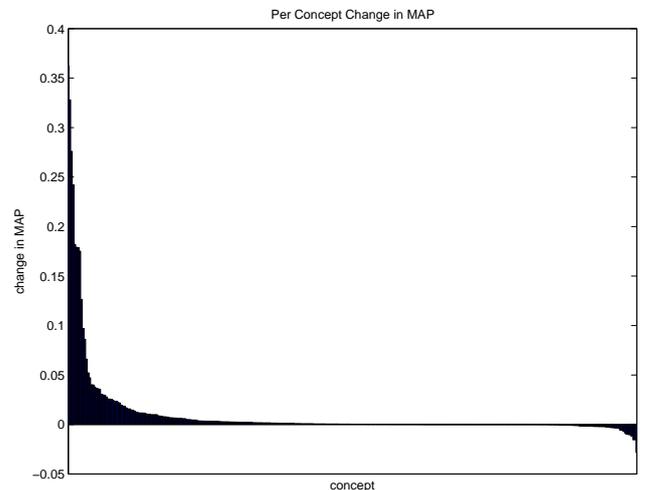
#### 5.5 Impact at the Item Level

Finally, we turn to an item-level discussion of the results. Figures 10 and 11 show an overview, for topics and concepts respectively, of the per-concept change in MAP between the baseline and the best performing retrieval method at window size 20. On the whole, when there is a change, most are positive. Taking into consideration changes of magnitude  $> 0.01$  only, 47% of topics and 14% of concepts are

<sup>4</sup>At window size 20.



**Figure 10: Per topic change in MAP score when comparing the optimal retrieval model at window size = 20 to the baseline, sorted by improvement over the baseline**



**Figure 11: Per concept change in MAP score when comparing the optimal retrieval model at window size = 20 to the baseline, sorted by improvement over the baseline**

positively affected while 6% and 2%, respectively, are negatively affected. The scale of change in the negative direction is much smaller ( $+0.55$  vs  $-0.09$  for topics, and  $+0.36$  vs  $-0.03$  for concepts).

Table 3 provides the text descriptions of items at the extreme ends of the plots in Figures 10 and 11.<sup>5</sup>

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we examined the redundancy phenomenon in video retrieval, i.e., the phenomenon that in news video

<sup>5</sup>Preliminary investigations indicate that the redundancy models tend to improve recall, by incorporating extra speech clues for each shot, while maintaining a relatively high precision by taking redundancy patterns into account.

**Table 3: Examples of topics and concepts affected by redundancy models**

Topics		
Item Id	$\Delta$ MAP	Text Description
0116	0.5542	the Sphinx
0114	0.3105	Osama Bin Laden
...	...	...
0109	-0.0747	one or more tanks
0106	-0.0864	the Tomb of the Unknown Soldier at Arlington National Cemetery

Concepts		
Item Id	$\Delta$ MAP	description
LSCOM292	0.3625	dog, domestic dog, Canis familiaris
MM007	0.3280	bicycle, bike, wheel, cycle
...	...	...
MM063	-0.0155	motorcycle, bike
LSCOM192	-0.0283	body, dead body

important information is often repeated several times, both within and across channels. By examining the data sets made available by TRECVID, we found redundancy patterns in the video channel describing the extent to which visual items cluster together, and between the video channel and the speech channel showing that speech is an indicator of visual relevance. In the latter case we also observed a temporal mismatch between topic/concept occurrences in the speech and visual channels. One of the main contributions of the paper in this respect has been to explore visual redundancy over many consecutive shots, and explore cross-channel redundancy for a large number of concepts and topics, arriving at a power law distribution (rather than a gaussian as proposed in the literature) to model the distribution.

Additional contributions concerned the incorporation of the redundancy phenomena that we uncovered into a retrieval framework—we considered both an empirical model (directly reflecting the data) and a power law model fitted to the data. These models were integrated into a language modeling-based approach to retrieval, through the use of document expansion. This way of incorporating redundancy phenomena into a retrieval framework led to a significant boost in retrieval performance. The improvements were achieved both on topics (that tend to have a somewhat elaborate textual description) and on concepts (with much more concise descriptions). And, importantly, the power law models of redundancy performed at least as well as the empirical distributions.

As to future work, this paper generalizes redundancy patterns over a large numbers of items, divided into topics and concepts. We intend to explore other item typologies and their effect on redundancy patterns. Another logical extension of this work is to investigate whether story boundaries can be used to further refine the retrieval models. Finally, this paper has concentrated on using redundancy phenomena to improve speech-based retrieval of visual items. It would be interesting to see whether these phenomena can be utilised in other multimedia analysis tasks.

## 7. ACKNOWLEDGMENTS

The authors would like to thank Jan-Mark Geusebroek for his valuable assistance in modelling the various distributions outlined in this paper.

Both authors were supported by the Netherlands Organization for Scientific Research (NWO) MuNCH project under project number 640.002.501. Maarten de Rijke was also supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612-000.106, 612.066.302, 612.069.006, 640.001.501, and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

## 8. REFERENCES

- [1] M. Campbell, A. Haubold, S. Ebadollahi, M. R. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, and L. Xie. IBM research TRECVID-2006 video retrieval system. In *TREC Video Retrieval Evaluation Proceedings*, 2006.
- [2] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, X. Dong, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 video search and high-level feature extraction. In *TREC Video Retrieval Evaluation Proceedings*, 2006.
- [3] T.-S. Chua, S.-Y. Neo, Y. Zheng, H.-K. Goh, Y. Xiao, S. Tang, and M. Zhao. TRECVID 2006 by NUS-I2R. In *TREC Video Retrieval Evaluation Proceedings*, 2006.
- [4] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and J. Yang. Multi-lingual broadcast news retrieval. In *TREC Video Retrieval Evaluation Proceedings*, 2006.
- [5] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.
- [6] B. Huurnink and M. de Rijke. The value of stories for speech-based video search. In *CIVR, Lecture Notes in Computer Science*, pages 266–271. Springer, 2007.
- [7] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
- [8] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.
- [9] M. F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.
- [10] A. Singhal and F. Pereira. Document expansion for speech retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 34–41, New York, NY, USA, 1999. ACM Press.
- [11] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

- [12] S. W. Smoliar and H. Zhang. Content-based video indexing and retrieval. *IEEE MultiMedia*, 1(2):62–72, 1994.
- [13] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.
- [14] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9(5), August 2007. *In press*.
- [15] C. G. M. Snoek, J. C. van Gemert, T. Gevers, B. Huurnink, D. C. Koelma, M. V. Liempt, O. D. Rooij, K. E. A. van de Sande, F. J. Seinstra, A. W. M. Smeulders, A. H. Thean, C. J. Veenman, and M. Worring. The MediaMill TRECVID 2006 semantic video search engine. In *TREC Video Retrieval Evaluation Proceedings*, 2006.
- [16] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In R. C. Moore, J. A. Bilmes, J. Chu-Carroll, and M. Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006.
- [17] J. C. van Gemert, C. G. M. Snoek, C. J. Veenman, and A. W. M. Smeulders. The influence of cross-validation on video classification performance. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 695–698, New York, NY, USA, 2006. ACM Press.
- [18] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 33–42, New York, NY, USA, 2006. ACM Press.
- [19] J. Yang, M. yu Chen, and A. G. Hauptmann. Finding person X: Correlating names with visual appearances. In *CIVR*, volume 3115 of *Lecture Notes in Computer Science*, pages 270–278. Springer, 2004.
- [20] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.