

File ID 191357  
Filename Chapter 9: Redundancy and the temporal mismatch

---

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation  
Title Search in audiovisual broadcast archives  
Author B. Huurnink  
Faculty Faculty of Science  
Year 2010  
Pages viii, 198  
ISBN 978-90-786-7599-0

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/358972>

---

*Copyright*

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use.*

---

# Redundancy and the Temporal Mismatch

Having addressed an aspect of detector-based search in the previous chapter, we will move on in this chapter to explore a problem faced when searching on the basis of automatically generated speech transcripts. In Section 3.3.4 we noted that user keyword searches cover a wide range of words, with the majority of query terms occurring only a few times. Transcript-based search allows us to directly perform retrieval on a wide range of words, and potentially to find visual content. This suggests that transcript-based search may be well-suited to the needs of audiovisual broadcast archives. However, transcript-based search is sensitive to the *temporal mismatch*; a lack of alignment between the spoken mention of a word in the audio signal and its appearance in the video signal. Here we will address the problem of the temporal mismatch by modeling *redundancy* phenomena.

Redundancy is the phenomenon that narrative structure is used to repeat information that is important to a video across multiple shots, and in multiple ways [172]. For example, a person appearing in the video signal may also be mentioned by name in the dialog captured in the audio signal. In addition, the person may appear in multiple shots in close temporal proximity, especially if they are a central figure in the narrative. In this way, if we are temporarily distracted from listening to the dialog or looking at the screen, we still have an idea of the semantic content of the video. As we watch a video segment, we gain understanding of its content by integrating different forms of information over time.

Our working assumption is that redundancy across audio and video signals can be used to improve the effectiveness of video retrieval. In this chapter we are espe-

Time	Video signal	Audio signal (transcripts)	Visual match?	Transcript match?
1		he also said that have spread frank with the iraqi issue, he and <b>blair</b> has different chirac said he is wrong to pay tax at the final will focus on the prime minister <b>tony blair</b> more stress that the two countries to enforce our hope that the two countries hope to see a stable and democratic iraq the two countries support iraq	✗	✓
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' leaders think that is to support the iraqi government forces a significant new opportunities chirac's	✓	✗
3		arafat after passing the palestinian-israeli situation <b>blair</b> applying expressed the hope that palestine and israel to seize the opportunity reopen peace process music of the palestinian elections can be elected by the majority of the two countries' leaders expressed the hope that the fragrance of	✗	✓
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 in the state banquet	✓	✗

**Figure 9.1:** Finding Tony Blair: an example of temporal item distribution across the audio and video signals within series of shots from a news broadcast. Tony Blair is seen in the video signal of the second and third shots, while his name appears in transcripts (derived from the audio signal) of the first and third shots.

cially interested in using redundancy—both within the video signal, and across the video and audio signals—to improve the effectiveness of transcript-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, if a given shot is visually relevant, how likely is it that a neighboring shot is visually relevant as well? And when visual relevance does spread out to neighboring shots, at which distance can we still observe the effect? The phenomenon is illustrated in Figure 9.1. 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 across the audio and visual signals. When objects are difficult to detect on the basis of the video signal, a retrieval engine can look for clues in information captured from the audio signal. For example, individual people such as Tony Blair or Saddam Hussein tend to be

difficult for an automatic system to detect using only visual information. However, a news broadcast showing Tony Blair is likely to mention his name several times, and (depending on the performance of the automatic speech recognition system used) his name will therefore be present in the transcripts. For retrieval, an interesting challenge emerges here: taking into account the *temporal mismatch* that can occur across the audio and the video signals. While each signal is temporally cohesive in its own right, the content may not be synchronized across them. For example, in Figure 9.1 we see a displacement across the mention of Tony Blair’s name in shot 1, and his visual appearance in shot 2. It has been shown that named people are on average mentioned two seconds before they appear in broadcast news [195]. We perform an investigation into the distribution of visual to visual relevance (i.e., how likely it is for visual items to occur closely together) and contrast this with the distribution of cross-signal transcript to visual relevance (i.e., how likely it is for items to occur in the visual signal, at given temporal proximity to their mention in the audio signal as captured in transcripts).

Now, assuming that we have developed sufficient understanding of redundancy to model a distribution of the phenomenon, the next step is to try and exploit the phenomenon to improve the effectiveness of transcript-based search for visual items. In a transcript-based search 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 9.1, if transcripts indicate that shot 4 is relevant, there is an increased probability that surrounding shots are also relevant. In this study we find consistent redundancy patterns within and across the video and audio signals, and we propose retrieval models that integrate these patterns to improve cross-signal retrieval performance.

We center on the question,

**RQ 4** *Within a video broadcast, the same object may appear multiple times within and across the video and audio signal, for example being mentioned in speech and then appearing in the visual signal. How can this phenomenon be characterized, and can we model and use this characteristic so as to improve cross-stream retrieval of visual items using transcripts?*

More specific questions are:

**CRQ 1** Given a set of visual items in the form of queries or concepts, how are visually relevant shots distributed across shots in the video signal?

**CRQ 2** How can we characterize the temporal mismatch for visual items across the video and the audio signal?

**CRQ 3** How consistent are our characterizations between collections?

**CRQ 4** What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?

We address our questions through an empirical exploration of real-world video data and by developing a retrieval framework for incorporating temporal redundancy distributions in transcript-based search.

The remainder of the chapter is organised as follows. Section 9.1 is devoted to characterizing the redundancy phenomenon. Section 9.2 outlines the retrieval framework in which we incorporate redundancy models, and in Section 9.3 we describe the retrieval results and analysis. Conclusions are presented in Section 9.4.

## 9.1 Discovering Redundancy

Here we describe our studies into the characterization of redundancy and the temporal mismatch within and across the audio and video signals in audiovisual material.

### 9.1.1 Methodology

In order to answer CRQ 1 (about the temporal distribution of visually relevant items in the video signal), 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 [194], who are interested in the *transitional probability* of a neighbouring 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 estimated by  $\hat{p}(e_n = V_R | d = V_R)$ , 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 manually created judgments of  $V_R$  to calculate the transitional probability of an item at offset  $n$  according to

$$\hat{p}(e_n = V_R | d = V_R) = \frac{c(e_n = V_R, d = V_R)}{c(d = V_R)}, \quad (9.1)$$

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 there is no shot at offset  $n$ , due to the offset being outside the beginning or end of the video, then  $e_n \neq V_R$ .

To answer CRQ 2 (about the temporal mismatch across the audio and visual signals) we calculate the transitional probability of  $e_n$  given that  $d$  has a match in

**Table 9.1:** Overview of the collections and visual items used in the experiments in this chapter. Mean numbers of assessments of  $V_R$  and  $V_T$  per visual item are indicated by  $\bar{x}(V_R)$  and  $\bar{x}(T_R)$  respectively. Note that for queries  $\bar{x}(V_R) < \bar{x}(T_R)$ , while for concepts  $\bar{x}(V_R) > \bar{x}(T_R)$ .

Collection	Item Type	Item Statistics		
		# items	$\bar{x}(V_R)$	$\bar{x}(T_R)$
Broadcast News	Query	67	47	874
Broadcast News	Concept	450	1,189	201
Broadcast News+	Query	96	203	889
Archive Footage	Query	101	252	933

the *transcript* stream. Substituting the transcript relevance,  $T_R$  of  $d$  into Eq. 9.1 this gives

$$\hat{p}(e_n = V_R | d = T_R) = \frac{c(e_n = V_R, d = T_R)}{c(d = T_R)}, \quad (9.2)$$

where we say that  $d = T_R$  when the transcript associated with the shot contains one of the words of the item description.

## 9.1.2 Experimental Setup

**Collections** For our experiments we require a collection of shots that can be associated with shot-level assessments of  $V_R$  and  $T_R$  for different queries. An overview of the collections in terms of the numbers of queries and assessments of  $V_R$  and  $T_R$  is given in Table 9.1.

Both of the collections described in Chapter 7 provide us with such data; each collection contains queries and manually created judgments of  $V_R$ , and each shot is associated with automatically created transcripts of speech that can be used to assess  $T_R$ . In addition to queries, the Broadcast News collection contains judgments for the 450 visual concepts contained in the multimedia thesaurus; we use these concepts as an additional set of visual items.

Not all of the query sets in the Archive Footage collection are suited for characterizing redundancy. Specifically, we exclude the Archive query set from our experiments, as it not associated with manually created judgments of  $V_R$ ; the judgements for these queries were created using implicit data obtained from user purchases, and therefore video segments consisting of multiple shots, and sometimes entire programs, are grouped together (see Section 7.2.2). Therefore we exclude these from our experiments in this chapter, giving a combined set of 101 queries for the Archive Footage collection.

In order to provide additional data for evaluation and comparison, we define one

extra collection for our experiments, which we term the *Broadcast News+* collection. This collection is derived from the data sets used for TRECVID 2003 – 2006 benchmarking evaluations. Like the Broadcast News collection, the video data in this collection consists of English, Chinese, and Arabic news broadcasts, and is accompanied by queries, manually created relevance judgments, and (machine translated) English language transcripts. It differs to the Broadcast News collection in that it is not annotated with respect to the multimedia thesaurus of 450 concepts. In total this collection contains over 190,000 shots, and is associated with 96 textual queries as outlined in Table 9.1.

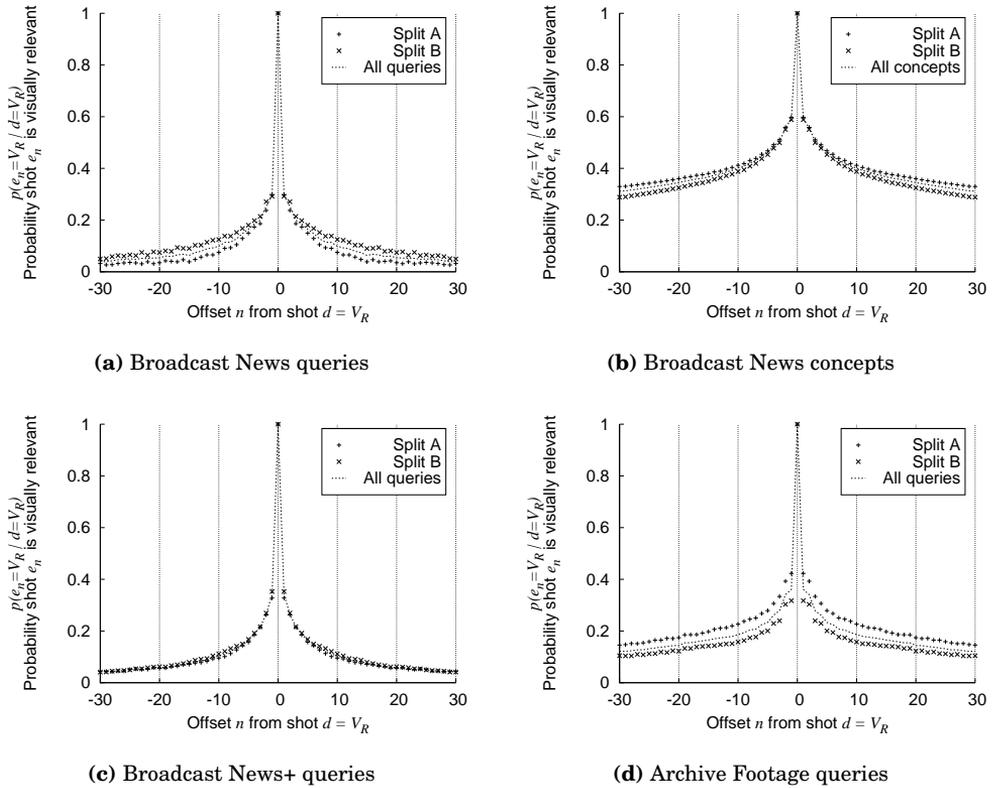
**Assessing  $T_R$**  The transcripts associated with each collection are used to assess  $T_R$  for the shots in the video data. This is done by matching text from the queries and concept descriptions to the text in the transcript of a shot. If a term from the transcript of  $d$  is matched for a query or concept, then  $d = T_R$ . Concept descriptions consist of synonyms (synsets) obtained from WordNet using the links from the unified multimedia thesaurus. Query descriptions consist of the natural language descriptions of information need described in Chapter 7.

**Partitioning visual items** We randomly split the visual items for each combination of collection and item type into two sets, *Split A* and *Split B*. We do this to avoid overfitting our models of redundancy patterns to our retrieval experiments. In our retrieval experiments we will use models developed on Split A when performing retrieval for items in Split B, and vice versa.

### 9.1.3 Redundancy in the Video Signal

Given that the current shot contains a visually relevant item, what is the probability that a neighboring shot is also visually relevant? Figure 9.2 gives an overview of the transitional probabilities, calculated using Equation 9.1 and averaged over all instances of  $d = V_R$ , for the queries and concepts contained in the Broadcast News, Broadcast News+, and Archive Footage collections. The graphs are centered around the known visually relevant shot  $d = V_R$  in the middle; along the X-axis we plot the distance from this shot as measured in terms of the number of shots.

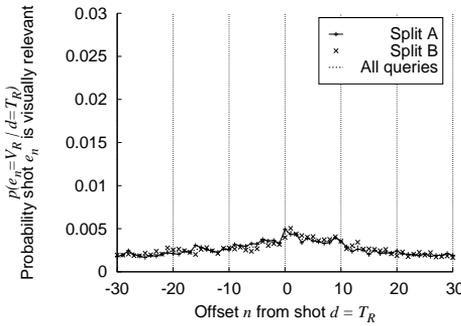
In Figure 9.2 concepts and queries are plotted separately for the different collections, and we see that the redundancy patterns exhibit a similar shape. They are all symmetrical, each pattern peaks sharply at the shot offset of 0 (the known visually relevant shot), and each pattern smooths out to the background probability that any random shot is visually relevant. These curves resemble a power law distribution, as we will discuss in Section 9.1.5. This contrasts to the Gaussian distribution observed by Yang et al. [195]; this because they defined the redundancy window in



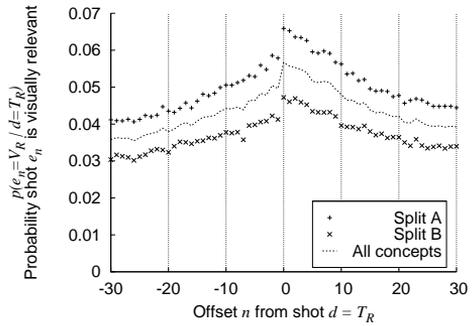
**Figure 9.2:** Visual redundancy patterns showing the transitional probability of a shot being visually relevant, based on its offset from a known visually relevant shot

terms of seconds, resulting in a smoother distribution. We, on the other hand, to enable a larger scale of analysis, have defined the redundancy windows in terms of shots. This choice was made because plentiful shot-level relevance data is available, however, shots are units that encompass uneven amounts of time, and this makes prevents the measurements from being evenly distributed.

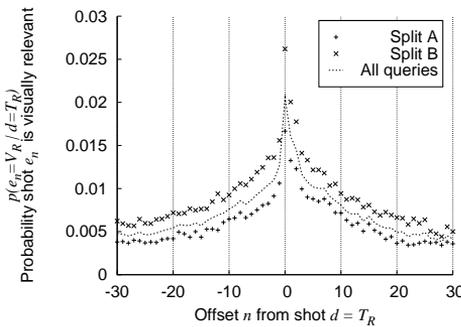
There is also a notable difference between the redundancy patterns, specifically between that for concepts in the Broadcast News collection shown in Figure 9.2b and the query distributions on the different collections shown in Figures 9.2a, 9.2c, and 9.2d. The concept pattern smooths out to a higher probability value than the query patterns do. This is because concepts have more visually relevant shots on average than queries do. For example, the concept with the most relevant shots in the Broadcast News collection, *Person*, occurs in 33,869 of the 43,907 shots, and has a background probability of 0.95. In contrast, the query with the most relevant shots



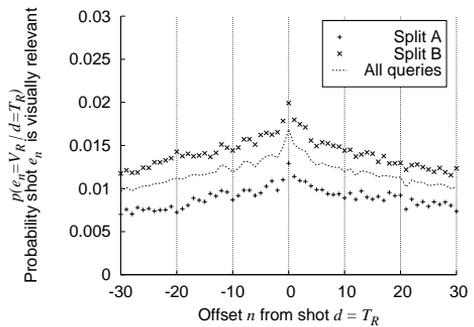
(a) Broadcast News queries



(b) Broadcast News concepts



(c) Broadcast News+ queries



(d) Archive Footage topic

**Figure 9.3:** Cross-signal redundancy patterns showing the transitional probability of a shot being visually relevant, based on its offset from a match in the transcripts. Note the difference in scale for the Broadcast News concepts, which on average have more relevant shots in the collection than do the queries.

in the Broadcast News collection, *Presidential candidates*, occurs in 240 shots and has a background probability of only 0.01. With concepts occurring more frequently than queries in the shots in the collection, a given concept is more likely to occur in a randomly selected shot than a given query.

### 9.1.4 Redundancy Across the Video and Audio Signals

Figure 9.3 shows the transitional probability of visual relevance for surrounding shots, given that the transcript of a shot  $d = T_R$  contains an item word. Note the scale differences compared to Figure 9.2, which demonstrate that  $T_R$  by no means directly indicates  $V_R$ ; when the transcript of a given shot contains a query or concept

**Table 9.2:** The average probability that a neighbouring shot is visually relevant, for the 30 shots before the text match and the 30 shots after the text match. The difference in the average probability before and after the text match is indicated by  $\Delta$ . In collections of broadcast news, visually relevant shots are more likely to occur after a text match than they are to occur before a text match.

Collection	Item Type	Average probability $e_n = V_R$		
		$n \in [-5 .. -1]$	$n \in [1 .. 5]$	$\Delta$
Broadcast News	Query	0.0025	0.0028	10%
Broadcast News	Concept	0.0413	0.0454	10%
Broadcast News+	Query	0.0068	0.0072	6%
Archive Footage	Query	0.0118	0.0118	0%

word, the probability that the shot is visually relevant is still very low. Though the scale has changed, we see evidence of redundancy across the signal sources for both queries and concepts: there is a peak in probability of visual relevance close to the point where a transcript match occurs. This is most pronounced for the Broadcast News+ queries.

Furthermore, unlike redundancy within the visual signal, the distributions here are not symmetrical around the central shot  $d$ . There is evidence of a temporal mismatch, with shots being more likely to be visually relevant after the transcript match than they are before the match. This is especially apparent in Figure 9.3b, where the curve after  $d = T_R$  is less steep than the curve after it. We quantify the temporal mismatch effect over the first 30 shots before and after the text match in Table 9.2. For all the item types in the collections based on broadcast news, a visually relevant shot is more likely to appear after the text match than before the text match. For the queries in the Archive Footage collection, a visually relevant shot is equally likely to appear before the text match as it is to appear afterwards.

We have now uncovered redundancy phenomena and a temporal mismatch across the video and the audio signals in video; next we will model these empirical phenomena, so that we can integrate them into a retrieval system.

### 9.1.5 Estimating Expected Visual Relevance

This section is devoted to estimating the expected visual relevance  $e_n$  for the shots surrounding  $d$ . We will estimate  $\gamma$  in two ways: (1) by the empirically derived distribution, and (2) by a simplified approximation of the distribution.

To estimate  $\gamma$  from empirical redundancy patterns, we scale each pattern 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 collection.

To estimate  $\gamma$  by an approximation of the empirical redundancy patterns, we

build on the observation made previously that the visual redundancy patterns resemble power law distributions. We use logistic regression on the data points to develop power law functions 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-signal redundancy patterns, 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. This power law will be used to inform the retrieval model that we develop in the next section.

## 9.2 Retrieval Framework

Our exploration of redundancy has shown that the visual content of a shot is to some extent reflected in the transcript of surrounding shots. Therefore, we wish to adjust the transcript of each shot with transcript text from the surrounding shot neighbourhood, and to examine the effect of incorporating our estimations of visual relevance into retrieval. To do this we must develop a retrieval framework for transcript-based search.

### 9.2.1 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 [60, 122, 197]. Furthermore, the philosophy behind language modeling fits well with our retrieval wishes. Let us explain.

Under the standard language modeling approach, we assume that a document  $d$  is generated by a random sample of unigrams from a hidden document model  $\theta_d$ , where  $\theta_d$  is a document-specific probability distribution [96]. At retrieval time, for a query  $q$ , each document is ranked with respect to the probability that  $q$  was generated by  $\theta_d$ . Therefore, the essential problem is estimating  $\theta_d$ . Assuming an estimated model  $\hat{\theta}_d$  of document  $d$  and a query  $q$  containing words  $w_1, w_2, \dots, w_m$ , we rank  $d$  according to  $p(q|\hat{\theta}_d)$  so that

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

One approach to determining  $p(w|\hat{\theta}_d)$  is to use maximum likelihood estimation (MLE). A simple MLE estimate of  $p(w|\hat{\theta}_d)$  is 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 doc-

ument 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 [197], as we have previously found this to be a suited method for transcript-based video retrieval [71]. This method interpolates the maximum likelihood with the background collection language model  $\theta_C$ . The Jelinek-Mercer smoothing estimate is given by

$$p(w|\hat{\theta}_d) = \lambda \cdot \frac{c(w, d)}{|d|} + (1 - \lambda) \cdot p(w|\theta_C), \quad (9.3)$$

where  $\lambda$  is a fixed parameter that controls the interpolation.

Redundancy will be integrated within our retrieval model by adjusting the word counts  $c(w, d)$ , as we will now explain. 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, documents are shots associated with transcripts. We use redundancy information to help estimate  $\theta_d$ , and adjust the word counts for a document with transcripts from surrounding shots to help estimate  $\theta_d$  for each shot.

## 9.2.2 Document Expansion

Document expansion is a technique originating from spoken document retrieval that allows for incorporation of external evidence in a natural way [140]. 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. Tao et al. [164] propose a general model for document expansion in the language modelling setting, on which we build. To perform document expansion, we use a set of external shots  $E$  to determine additional information about every shot  $d$  in a collection  $C$ . At indexing time we use word counts from the transcripts associated with  $d$  and from transcripts associated with the shots in  $E$  to create a transcript associated with a ‘pseudo-shot,’  $d'$ . The word counts in  $d'$ ,  $c(w, d')$ , are adjusted from those in  $d$  according to:

$$c(w, d') = \alpha \cdot c(w, d) + (1 - \alpha) \cdot \sum_{e \in E} (\gamma_d(e) \cdot c(w, e)), \quad (9.4)$$

where  $\alpha$  is a constant,  $e$  is a shot 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 the transcript of  $d$ .

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

- our central shot  $d$ , and its associated transcript;
- our set of external shots  $E$ , which we define as the neighbouring shots within a window of  $n$  shots;
- $\gamma$  is our model of the expected visual relevance for  $e \in E$ .
- as  $d$  is always the central member of  $E$ , this eliminates the need for  $\alpha$ , which is replaced by the  $\gamma$  value at offset 0.

This leads to the following temporally expanded document model:

$$c(w, d') = \sum_{e \in E} (\gamma_d(e) \cdot c(w, e)), \quad (9.5)$$

We arrive at different retrieval models by making different choices for  $\gamma$ ; then, Eq. 9.5 is used instead of the original word count  $c(w, d)$  in Eq. 9.3.

### 9.2.3 Integrating Redundancy

Finally, we need to put together the two main ingredients developed so far: our retrieval framework and the models of expected visual relevance described in Section 9.1.5.

We integrate redundancy into our framework by modifying the  $\gamma$  function in Eq. 9.5. We consider two baselines; for the first no document expansion is performed at all, and for the second document expansion is without incorporating any information about the expected visual relevance of neighbouring shots. We develop four retrieval models based on the different models of expected visual relevance described in Section 9.1.5. The retrieval experiments are denoted as follows:

**B1. No expansion.** No integrated redundancy; use the model described in Eq. 9.3;

**B2. Flat.**  $\gamma = 1$ : all shots are expected to be equally visually relevant;

**D1. Visual data driven.**  $\gamma$  is determined by the empirical visual redundancy value, at distance  $n$ ;

**D2. Transcript data driven.**  $\gamma$  is determined by the empirical cross-signal redundancy value, at distance  $n$ ;

**P1. Visual model driven.**  $\gamma$  is determined by a power law approximation of visual redundancy, at distance  $n$ .

**P2. Transcript model driven.**  $\gamma$  is determined by a power law approximation of cross-signal redundancy, at distance  $n$ ;

**Table 9.3:** MAP scores for the different retrieval models at window size 30. The highest performing score for each combination of item type and collection is indicated in bold.  $\blacktriangle$ ,  $\blacktriangledown$ , and  $\circ$ , respectively indicate that a score is significantly better than, worse than, or statistically indistinguishable from the scores of B1 and B2, from left to right.

Collection	Item type	Retrieval model					
		Baselines		Data driven		Power law	
		B1	B2	D1	D2	P1	P2
Broadcast News	Query	0.006	0.005	0.010 $\blacktriangle\blacktriangle$	0.008 $\circ\blacktriangle$	<b>0.011</b> $\blacktriangle\blacktriangle$	0.008 $\circ\blacktriangle$
Broadcast News	Concept	0.005	0.011	<b>0.013</b> $\blacktriangle\blacktriangle$	0.012 $\blacktriangle\blacktriangle$	<b>0.013</b> $\blacktriangle\blacktriangle$	<b>0.013</b> $\blacktriangle\blacktriangle$
Broadcast News+	Query	0.040	0.030	<b>0.076</b> $\blacktriangle\blacktriangle$	0.069 $\blacktriangle\blacktriangle$	0.073 $\blacktriangle\blacktriangle$	0.069 $\blacktriangle\blacktriangle$
Archive Footage	Query	0.006	0.021	0.022 $\blacktriangle\circ$	0.023 $\blacktriangle\circ$	0.020 $\blacktriangle\circ$	<b>0.025</b> $\blacktriangle\circ$

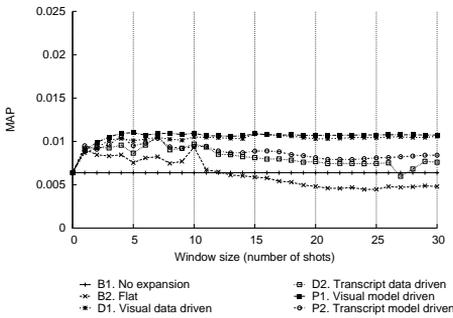
## 9.3 Retrieval Experiments

In our retrieval experiments we use the retrieval framework developed in Section 9.2 to address CRQ 4, *What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?* We test each of the retrieval models described in Section 9.2.3 using the data sets and items described in Section 9.1. The models are evaluated at increasing shot window sizes.

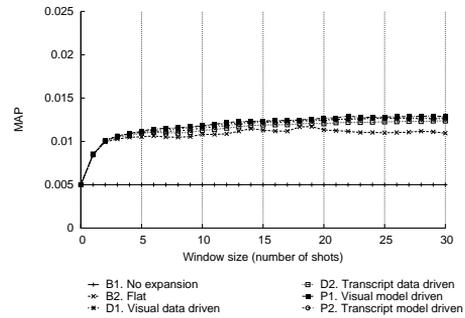
### 9.3.1 Result Overview

Figure 9.4 provides an overview of the retrieval results for each of the retrieval models that we consider across increasing shot window sizes, while Table 9.3 specifies in numbers the retrieval scores of the retrieval models, and significant differences to the baseline, at a window size of 30 shots. In general we observe that the MAP scores initially increase as the window size increases, indicating that using document expansion always increases overall performance when retrieving visual items using transcripts. However, as transcripts from increasing numbers of shots are incorporated in the expanded documents, differences between retrieval models become clear, as we will discuss for each combination of collection and item type in turn.

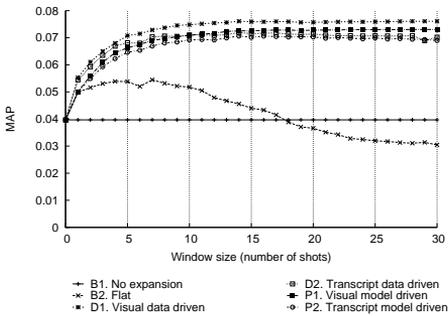
First, let us examine the performance of the retrieval models on the queries in the Broadcast News collection, shown in Figure 9.4a. Here the two models based on visual redundancy, D1 and P1, attain the highest performance, leveling off at a window size of between 5–10 shots. The two models based on cross-signal redundancy patterns, D2 and P2, exhibit different behavior — rather than leveling off as D1 and P1 do, performance starts to degrade at a window size of 5 shots. D2 and P2 do not perform significantly better than B1 at a window size of 30 shots, though they do perform significantly better than B2, which also degrades as more shots are added to the expanded documents.



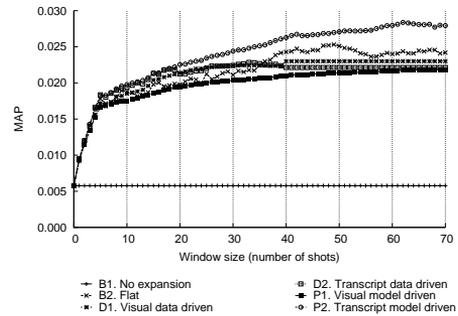
(a) Query MAP scores on the Broadcast News collection



(b) Concept MAP scores on the Broadcast News collection



(c) Query MAP scores on the Broadcast News+ collection



(d) Query MAP scores on the Archive Footage collection

**Figure 9.4:** Results of retrieval experiments across increasing window sizes. At a window size of 30, the transcript of each shot is expanded with the text from the transcripts of the 30 preceding shots and the 30 subsequent shots. The retrieval performance of the queries on the Audiovisual Archive collection keeps rising to a window size of 62 shots.

The performance of the retrieval models on the concepts in the Broadcast News collection exhibits a different pattern, as we can see in Figure 9.4b. Here all of the redundancy based retrieval models have similar retrieval performance as window size increases, leveling off at a window size of about 20 shots. Significance tests at a window size of 30 shots showed that there is no significant difference between the scores of D1, D2, P1, and P2. B2 similarly increases in retrieval performance, but reaches a lower maximum than the four retrieval models based on redundancy patterns, which significantly improve upon it.

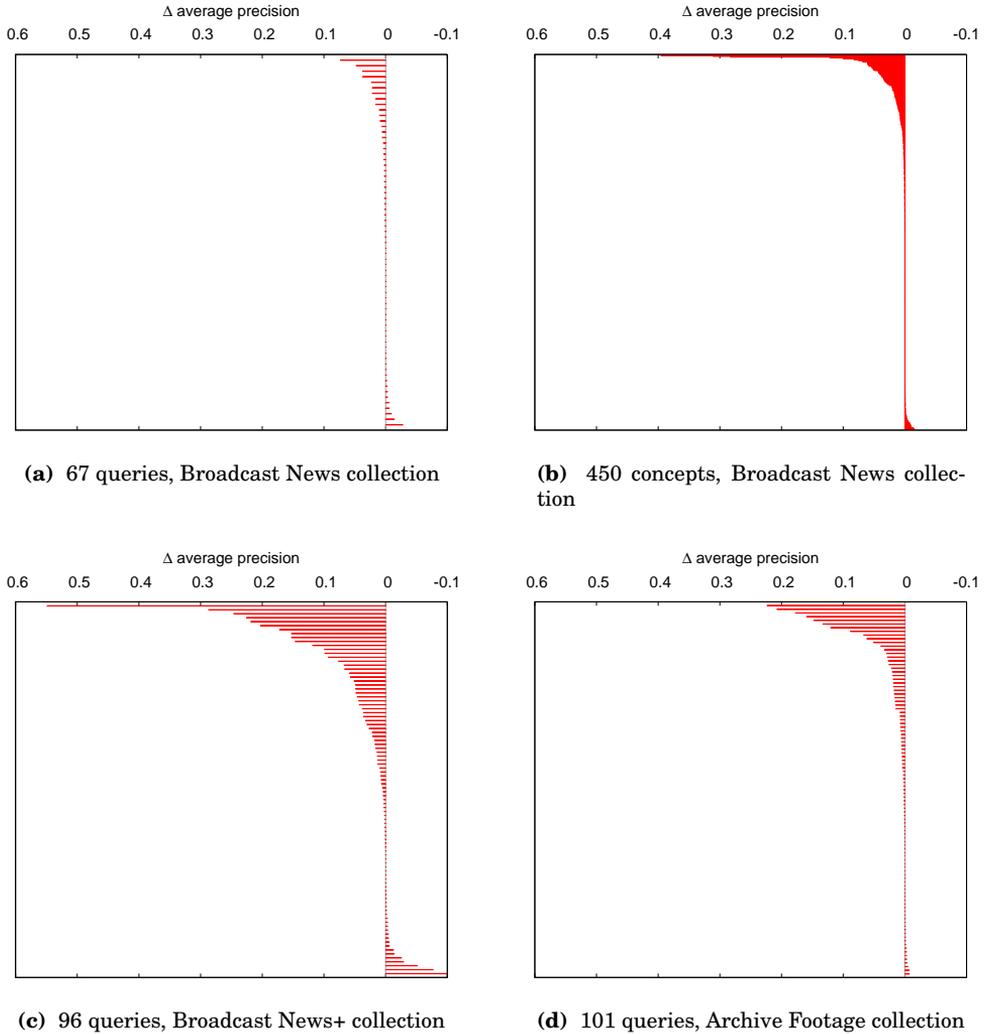
Differences between retrieval models based on redundancy and the two baselines are the greatest, in terms of absolute MAP scores, for the queries in the Broadcast News+ collection. Absolute retrieval performance here is so much higher than that of

the remaining visual items and collections that, for visualization purposes, the scale of Figure 9.4c had to be altered as compared to the remaining figures. Once again, the performance of all four models based on redundancy patterns increases steadily as window size increases, leveling off at a window size of about 15 shots. The two best performing models, D1 and P1, are based on visual redundancy patterns. Of the two, D1, which is based on empirical data measurements, outperforms P2, which is based on power law approximations of those measurements (though not significantly so). Baseline B2 starts to taper off rapidly when using a window size of more than 7 shots, and D1, D2, P1, and P2 all significantly outperform both baselines at a window size of 30 shots.

Finally, we turn to the performance of our retrieval models on queries from the Archive Footage collection, shown in Figure 9.4d. The performance here is unusual in that it is still increasing for all models at the maximum shown window size of 30. For this collection we performed retrieval experiments with windows of up to 70 shots, and found that retrieval performance levels out at a distance of 62 shots, with a MAP score of 0.028 for the best performing retrieval model, P2. At this point P2 is significantly better than both baselines. As for baseline B2, this also improves steadily as window size increases, though its performance is significantly worse than P2 at window size 30.

### 9.3.2 Impact at the Item Level

We turn to a discussion of responsiveness of individual items in our collections to redundancy-based models. Figure 9.5 gives an item-level specification of the change in average precision when comparing results from the best performing retrieval model to results from our first baseline, B1, which uses no document expansion at all, at window size 30. From this figure we can see that the majority of visual items benefit from incorporating document expansion. Taking into consideration all visual items across all collections, and taking into consideration changes of magnitude  $> 0.01$  only, retrieval performance is improved for 21% of the visual items, while it degrades for only 2% of the visual items. The largest absolute changes in performance can be observed for the queries in the Broadcast News+ collection. The query that improves the most ( $\Delta = 0.549$ ) is *Find shots of the Sphinx*. A manual inspection of the results revealed that there is one news story discussing the Sphinx in the collection, and it shows 12 visually relevant shots within a window of 15. At the same time the word “Sphinx” occurs only three times in the transcripts of the shots in this story, so by expanding the transcripts the relevant shots are retrieved and performance increases. The query that decreases most in performance ( $\Delta = -0.104$ ) is *Find shots of the Tomb of the Unknown Soldier at Arlington National Cemetery*. A manual inspection of these results showed that the 31 visually relevant shots for



**Figure 9.5:** Increase  $\Delta$  in average precision when using the optimal retrieval model at window size = 30 as compared to baseline B1, for the visual items in the Broadcast News, Broadcast News+, and Archive Footage collections. The visual items on the Y axis are sorted by increase over the baseline.

this query are distributed across six different news stories in the collection, and appear in tight clusters. Of the visually relevant shots, 17 contain at least one of the query words and are given a high ranking before expansion. Document expansion introduces irrelevant shots, decreasing the ranking of the relevant results.

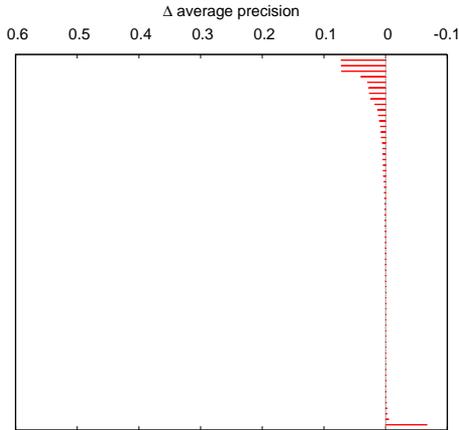
Figure 9.6 gives a specification of the change in average precision when comparing results from the best performing retrieval model to the second baseline, B2. This allows us to compare our redundancy models to a model that does include document expansion, but that does not take redundancy patterns into account. Once again, the majority of items benefit from a retrieval model that includes redundancy patterns. A total of 14% of the items are positively affected, and 3% of the items are negatively effected. The largest increase in performance ( $\Delta = 0.398$ ) is observed for a query from the Broadcast News collection, *Find shots of Boris Yeltsin*. The largest decrease in performance ( $\Delta = -0.094$ ) is for a query from the Archive Footage collection, *Find shots of apes or monkeys*. This query performs poorly as compared to flat expansion because most of the relevant shots are contained in a single documentary program about zoos. These shots are distributed throughout the program as background footage, while monkeys are only mentioned a few times. Therefore a flat window approach performs better here.

### 9.3.3 Comparing the redundancy models

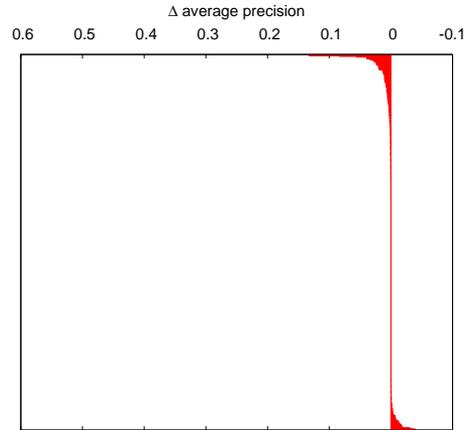
From our experiments it is clear that transcript-based search benefits in all cases from incorporating a retrieval model that takes redundancy into account. Now we turn to an examination of the differences across the individual retrieval models based on the redundancy patterns.

**Visual redundancy vs cross-signal redundancy** First we discuss the differences between the performance of retrieval models based on visual redundancy patterns (D1 and P1) as compared to the performance of retrieval models based on cross-signal redundancy patterns.

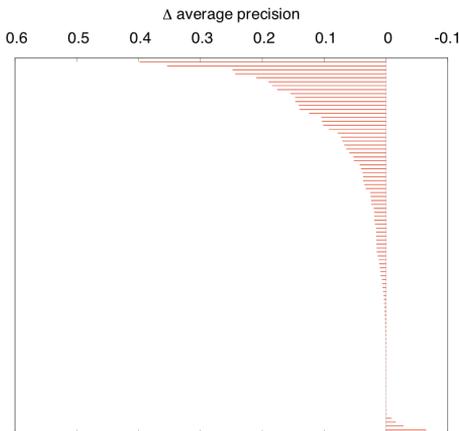
In two of the four cases, a redundancy model based on visual redundancy outperformed a model based on cross-signal redundancy. We found this surprising, as the cross-signal redundancy patterns assign shots occurring after the mention of a word in the transcript a higher estimation of visual relevance than shots that occur before the mention. Models using visual redundancy patterns, on the other hand, assign estimations of visual relevance symmetrically around  $d$ . So why did the queries in the Broadcast News and Broadcast News+ collections not respond as well to the models based on cross-signal redundancy? Turning back to the visual representations of the redundancy patterns in Figure 9.3d, we see that for the Broadcast News queries the visual redundancy patterns between splits do not agree. For Split A shot  $d$  has the maximum probability of being visually relevant, while for Split B shot  $d+1$  has the maximum probability of being visually relevant. For retrieval the splits are switched, causing incorrect estimations of visual relevance using the cross-signal redundancy models. As for the queries in the Broadcast News+ collection, we can



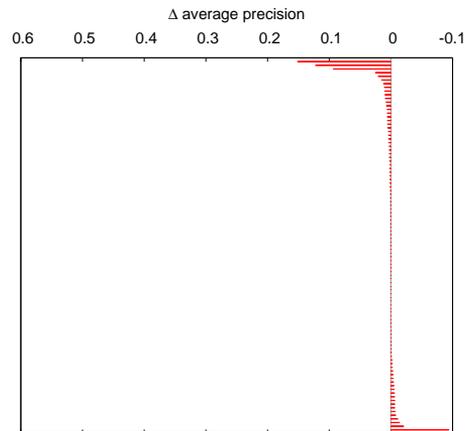
(a) 67 queries, Broadcast News collection



(b) 450 concepts, Broadcast News collection



(c) 96 queries, Broadcast News+ collection



(d) 101 queries, Archive Footage collection

**Figure 9.6:** Increase  $\Delta$  in average precision when using the optimal retrieval model at window size = 30 as compared to baseline B2, for the visual items in the Broadcast News, Broadcast News+, and Archive Footage collections. The visual items on the Y axis are sorted by increase over the baseline.

see from Table 9.2 that there is no strong temporal mismatch phenomenon for these queries in this collection, with a visually relevant shot being only 1% more likely to occur in the five shots after  $d$  than the five shots before  $d$ .

**Empirical measurements vs power law approximations** Turning to the differences between redundancy models that use empirical measurements (D1 and D2) and redundancy models that use power law approximations (P1 and P2), there is no significant difference between the two kinds of model. Therefore power law approximations may be used to stand in for empirical measurements when applying redundancy models in our framework.

## 9.4 Conclusions and Future Work

In this chapter we studied the redundancy phenomenon in video retrieval, in a broadcast news setting and audiovisual archive footage, i.e., the phenomenon that in the narrative of video important information is repeated, both within the video signal, and across the audio and video signals. We formulated four chapter level research questions to direct our study.

In answer to CRQ 1, *Given a set of visual items in the form of queries or concepts, how are visually relevant shots distributed across shots in the video signal?*, we found that visual redundancy patterns across shots resemble a power law function; given that a shot is visually relevant to an item, the probability that a neighbouring shot is also visually relevant decreases rapidly as the distance between the two increases. Turning to CRQ 2, *How can we characterize the temporal mismatch for visual items across the video and the audio signal?*, we observed there is redundancy, in that when an item is mentioned in the transcript (derived from the audio signal) of a shot, it is more likely to appear in either that shot or a neighbouring shot than it is to appear in a distant shot. Furthermore, we observed a temporal mismatch, with an item more likely to appear immediately after it is mentioned in the transcript than it is to appear before the transcript. In answer to CRQ 3, *How consistent are our characterizations between collections?*, we measured this phenomenon across different collections, and found that the temporal mismatch effect was stronger in news broadcasts than it was in more heterogeneous video data obtained from archive footage.

Finally, in order to answer CRQ 4 *What is the effect on the performance of transcript-based search of incorporating characterizations of redundancy and the temporal mismatch?*, we developed a retrieval framework that allows us to incorporate redundancy phenomena in transcript-based search. In this framework, the transcript of a shot is expanded with the (weighted) text of transcripts from surrounding shots. We found that combining transcripts from surrounding shots improved overall retrieval performance, as compared to retrieval performance when search on text from only a single shot, for all our collections and item types. In addition we found that, for all of our collections and item types, retrieval models that expanded transcripts on

the basis of redundancy phenomena outperformed retrieval models that expanded transcripts without incorporating redundancy phenomena. For three out of four collections and item types the increases were significant.

In this chapter we observed that retrieval performance in terms of absolute scores was low. While transcripts can give us some indication of the presence of visual items in the video signal, they are not always effective. In the next chapter we place transcript-based search in the context of the archive of the future, which we postulate will include detector-based search, feature-based search, and transcript-based search, as well as search on the manually created annotations that are already present in the archive.