# Novelty Detection for Cross-Lingual News Stories with Visual Duplicates and Speech Transcripts

Xiao Wu[+#] wuxiao@cs.cityu.edu.hk
Alexander G. Hauptmann[#] alex@cs.cmu.edu
Chong-Wah Ngo[+] cwngo@cs.cityu.edu.hk

[+]Department of Computer Science
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong

[#]School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, USA

## ABSTRACT
An overwhelming volume of news videos from different channels and languages is available today, which demands automatic management of this abundant information. To effectively search, retrieve, browse and track cross-lingual news stories, a news story similarity measure plays a critical role in assessing the novelty and redundancy among them. In this paper, we explore the novelty and redundancy detection with visual duplicates and speech transcripts for cross-lingual news stories. News stories are represented by a sequence of keyframes in the visual track and a set of words extracted from speech transcript in the audio track. A major difference to pure text documents is that the number of keyframes in one story is relatively small compared to the number of words and there exist a large number of non-near-duplicate keyframes. These features make the behavior of similarity measures different compared to traditional textual collections. Furthermore, the textual features and visual features complement each other for news stories. They can be further combined to boost the performance. Experiments on the TRECVID-2005 cross-lingual news video corpus show that approaches on textual features and visual features demonstrate different performance, and measures on visual features are quite effective. Overall, the cosine distance on keyframes is still a robust measure. Language models built on visual features demonstrate promising performance. The fusion of textual and visual features improves overall performance.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering, Search process*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Video analysis,* I.5.3 [**Pattern Recognition**]: Clustering – *Similarity Measures;*

## General Terms
Algorithms, Design, Experimentation, Performance.

## Keywords
Similarity Measure, Novelty and Redundancy Detection, Language Model, Multimodality, Cross-Lingual Information Retrieval, Near-Duplicate Keyframes, News Videos


*MM'07,* September 23–28, 2007, Augsburg, Bavaria, Germany.


## 1. INTRODUCTION
Nowadays news videos are broadcast daily from different sources, languages, and countries. Especially with the explosion of the Internet, videos are easily accessible and their availability has grown exponentially. For example, users upload 65,000 video clips each day at video sharing website YouTube, and the daily video views are over 100 million [27]. Given the huge volumes of news videos, it becomes important to manage these videos in an automatic and efficient way. To facilitate effective search, retrieval, browsing and tracking of news stories, news story similarity measures play a critical role in measuring the novelty and redundancy among news stories. In this paper, we will explore the novelty/redundancy detection with visual duplicates and textual speech transcripts for cross-lingual news stories.

The scenario becomes complicated when considering news stories from multilingual sources such as Chinese, English and Arabic. Due to errors or lack of proper processing tools for speech recognition and machine translation, the text transcripts are usually noisy and even unavailable when dealing with multilingual videos. For instance, there are 42 videos without automatic speech recognition (ASR) or corresponding machine translation out of 230 videos in the official TRECVID 2005 cross-lingual news video corpus [26]. The situation is even more serious when the news videos are downloaded from the Web. In these situations, previous text-based approaches (e.g. [2], [33]) for detecting novelty and redundancy in news videos will be inapplicable.

Given the frequent absence of text transcripts, it is important to consider how visual information can be exploited. Different from traditional text-based news articles, news videos include both audio and visual channels. In addition to the text transcripts, news videos carry rich visual content. Following TRECVID, a news story is defined as a segment of a news broadcast with a coherent news focus, which is a meaningful and semantic unit about one topic. In broadcast videos, stories about the same topic are often accompanied by a number of shots that tend to be used repeatedly due to the lack of fresh footage materials or as an audience reminder. A statistic in [32] indicates that there are approximately 10%~20% repeated shots in news stories. For example, the scene of planes hitting the World Trade Center is repeatedly broadcast by stories of the "9/11 terrorist attack". These shots can be frequently reused with minor modifications either as a reminder of the story or due to a lack of video material in the current

footage. This provides the audience (and our system) with useful cues to measure the similarity of news stories.

Furthermore, textual and visual features complement each other for news stories. It is interesting and meaningful to exploit the interaction between textual and visual features when text transcripts are available. The employment of either textual or visual concepts alone may not be sufficient since either content can appear differently over time. Therefore, we will make use of the speech transcripts from the audio track and visual duplicates at the keyframe level from the visual track to assess the novelty and redundancy of news stories.

In this paper, we study the novelty and redundancy detection with the textual and visual features in the context of cross-lingual broadcast domain. The research questions that motivate this work are (1) the feasibility of measures built on visual information at the keyframe level for novelty and redundancy detection, (2) the effect of near-duplicate keyframe detection on visual language models and vector space model, (3) the sensitivity of smoothing methods and parameter selection for language models, (4) the performance of approaches built on visual features compared to text-based methods, (5) improvement in fusing textual and visual features.

The rest of this paper is organized as follows. In section 2 we give a brief description of related work. The proposed framework and approaches of novelty and redundancy detection are discussed in section 3. Section 4 presents the experiments. Finally, we conclude the paper with a summary.

# 2. RELATED WORK

## 2.1 Textual Information Retrieval

Text based similarity measures have been extensively studied previously for estimating the similarity of text passages. Novelty and redundancy detection has been mainly explored at three different levels: the event level [3], document level [6, 29, 33] and sentence level [2]. Novelty detection at the event level originated from the work of new event detection or first story detection [3] in Topic Detection and Tracking (TDT) [1] which investigated several aspects for the automatic organization of news stories in the textual area. The most related work to our research in novelty/redundancy detection at the document and sentence level are [2], [17] and [33]. However, these approaches are based purely on the textual concepts. Their performance, when applied to noisy news videos scenario is still unexplored.

In addition to the classic *tf-idf* document vector space approach, recent work in the last ten years has demonstrated the effectiveness of an alternative "language model" approach in which probabilistic models of text generation are constructed from documents. Language modeling methods have been successfully applied to speech recognition, machine translation, and natural language processing, which have attracted a lot of research attention due to its solid foundation in statistical theory. Thus, the language modeling framework was introduced to information retrieval and has performed well empirically [e.g. 22, 33]. The basic idea is to estimate a language model for each document and then rank documents by the likelihood of their language models. Furthermore, language models for multi-lingual tasks were explored in [13, 14]. However, to the best of our knowledge, there are few cases where the image or video information was implemented with language models [16, 24, 25], and it is still unclear whether language models are feasible and effective for the cross-lingual video novelty and redundancy detection.

## 2.2 Multimedia Information Retrieval

Shot similarity and clip similarity have been extensively addressed for retrieval and clustering. Previously, clip and video copy detection (e.g. [5, 11]) were investigated by using image similarity measure with low-level global features, for instance, color histograms. Fast signature-based methods (e.g. [5]) were proposed to identify similar clips, which use a global statistic of the low-level features in clips. Global signatures are suitable for matching clips with almost identical content. Clip similarity ranking [11] was built on top of shot similarity and combines temporal order, granularity and so on. Bipartite graph based algorithms were proposed in [21] to compare the similarity of two clips. However, shot similarity detection built on global features is not robust enough for news story similarity measure due to the complicated variations of keyframes [32]. Moreover, cross-lingual news story similarity measure remains a challenging problem that has seen little exploration.

*Near-Duplicate Keyframes* (NDK) provide critical cues for novelty/redundancy detection and topic threading. News videos provide richer information than text streams. In news videos, news stories are often accompanied by video shots that tend to be repeated during the course of the topic. Traditionally, a shot is usually represented by a keyframe. There are significant numbers of near-duplicate keyframes (shots) that exist in news stories. NDK are keyframes close to the exact duplicate of each other, but different in some capturing conditions (camera, camera parameter, view angle, etc.), acquisition times, rendering conditions or editing operations [32]. A couple of examples can be found in Figure 2. Some methods to detect near duplicate keyframes were proposed in [12, 20, 32]. Different from global features, local feature detection approaches based on interest points [12, 20] avoid the shortcoming of global features and achieve robust detection results. Recently, near-duplicate keyframes were also exploited in [4] to boost the performance of interactive video search. Hsu et al. [9] tracked four topics with visual duplicates and semantic concepts, and found that near-duplicates significantly improved the tracking performance. Zhu et al [35] presented a hierarchical video content description and summarization strategy supported by a joint textual and visual similarity. Zhai et al. [31] linked news stories by combining keyframe matching and textual correlation. In our early work [28], we combined the text and human labeled NDK to thread news topics. Furthermore, semantic concept detection (e.g. [18], [19], [23]) has been investigated by many researchers, which serves an intermediate step in bridging the semantic gap between the low-level features and high-level concepts. Usually multiply low level features are extracted and each feature is trained with a supervised detector. Different supervised detectors were fused to offer complementary prediction. However, until now there has been no comprehensive exploration of the novelty and redundancy detection for cross-lingual news stories, which motivates this research.

# 3. NOVELTY/REDUNDANCY DETECTION

In this section, we discuss different measures using the vector space model and language models to detect the novelty and redundancy across cross-lingual news stories. The framework is first introduced in section 3.1, and the visual features from keyframes are represented in section 3.2. The cosine distance and language models are described in section 3.3 and 3.4 respectively. These methods are applicable for both speech transcripts and visual duplicates. Finally they are further fused to improve the performance.

## 3.1 Framework

The framework of cross-lingual news story similarity detection is shown in Figure 1, of which the novelty and redundancy detection is the focus of this paper. News videos from multiple sources with different languages constitute the news database. Topics are formed by clustering algorithms or classifiers. News stories related to a specific topic are presented in a chronological order based on their publication time, and each one must be evaluated before the next is seen. For simplicity, we assume all news stories in a topic are relevant, and the earliest news story that appeared in this topic is the first story of the topic. News stories are meaningful units which include audio and visual contents related to events. A news story consists of a sequence of shots in the visual track, and a set of words extracted from speech transcripts in the audio track. If a news story is from another language besides English, its text transcripts are translated into English counterparts by machine translation. Usually, a representative keyframe is extracted to represent each shot, and thus a story can be viewed as a list of shots represented by keyframes in the visual stream. Near-duplicate keyframe detection is performed to check the visual similarity among keyframes. In this paper, we use robust matching method based on local points to detect the NDK [20]. If keyframes are also appeared in other places of the whole corpus, they will be regarded as near-duplicate keyframes (NDK). Otherwise, they are non-near-duplicate (Non-NDK). The textual and visual information (NDK, Non-NDK, and text) is sent to the novelty and redundancy detection module to evaluate its novelty by comparing it with previously broadcast news stories. In this paper, we mainly focus on the novelty and redundancy detection. The novelty (redundancy) of news stories is measured at textual and visual track with vector space model and language models respectively. For language models, different smoothing techniques are implemented. Finally, textual and visual features are further fused to improve the performance.

The redundancy of a new news story $S_i$ is computed through a pairwise comparison between $S_i$ and every previously seen news story $S_j$. The previously seen news story most similar to $S_i$ determines the redundancy ($R$) of $S_i$.

$$R(S_i \mid S_1,...,S_{i-1}) = \max_{1 \le j \le i-1} R(S_i \mid S_j)$$

We treat novelty and redundancy as opposite ends of a continuous scale. Therefore ranking the news stories by increasing redundancy score is equivalent to ranking them by decreasing novelty score [33].

## 3.2 Visual Information

Similar to documents that are treated as a bag of words, we can treat a news story as composed of a bag of keyframes in the visual

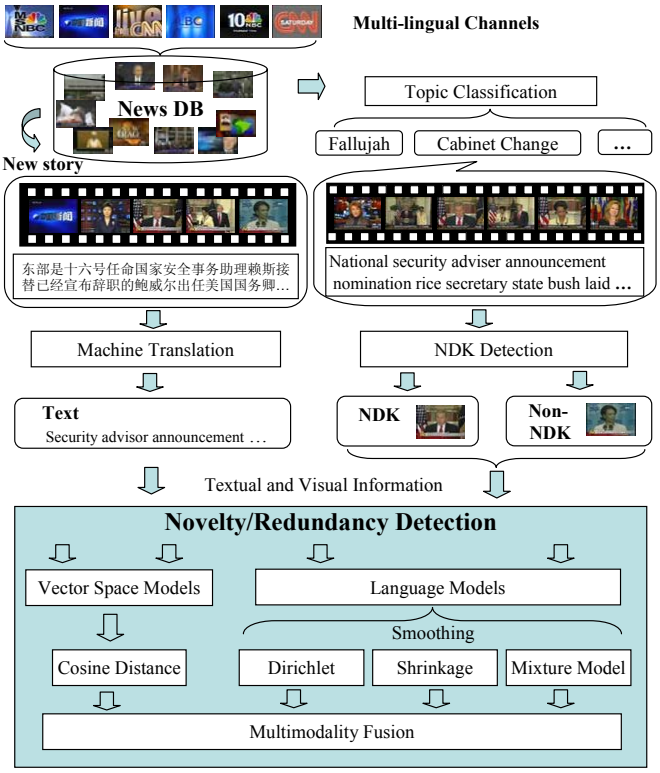



**Figure 1. Novelty and redundancy detection in multi-lingual multimedia environment (in which NDK means near-duplicate keyframes)**

track. The keyframes are further classified as near-duplicate keyframes (NDK) when they appear multiple times in the corpus and as non-near-duplicate keyframes (Non-NDK) when they appear only once. A keyframe can be treated as a special kind of words (that is: keyframe = visual keyword). Statistically, news stories are represented as smoothed probability distributions over the keyframes (visual keywords). The techniques used in text retrieval and classification can then be applied to the keyframes. We define the frequency of NDK in one news story as the visual term frequency (*tf*), and the frequency of NDK that appear in different news stories as the document frequency (*df*). For example, four sets of near-duplicate keyframes that appear in three news stories on "President Election" are shown in Figure 2 with different color borders. Among them, the first two stories are from English channels and the third is from a Chinese channel. The term frequency (*tf*) of the NDK with a red border in story $S_1$ (i.e. the keyframes that Bush is addressing) is two because the NDK appeared twice in story $S_1$. Its document frequency (*df*) is three since the NDK appeared in these three news stories. For Non-NDK, the visual term frequency and document frequency are both 1.

There are two major differences between keyframes (visual keywords) and traditional words: (1) the number of keyframes in each story is significantly smaller than the number of words, as each news story usually includes somewhere between 5 and 30 shots (2) there exists a very large number of Non-NDK that only appear once in the corpus. The impact of these differences has not been studied previously, particularly, with respect to their effects on document similarity measures.


**Figure 2. Near-duplicate keyframes appeared in three news stories of different channels (*tf: term frequency of keyframe, df: document frequency of keyframe*)**

## 3.3 Cosine Distance

The *Cosine distance* metric is a popular similarity measure in information retrieval. The cosine of the angle between a news story vector and each previously delivered news story vector determines the redundancy score for that news story. It is a pairwise measure, defined by

$$R(S_i, S_j) = \frac{\sum_{k=1}^{m} f_k(S_i) f_k(S_j)}{\sqrt{\sum_{k=1}^{m} f_k(S_i)^2 \sum_{k=1}^{m} f_k(S_j)^2}}$$

where $f_k(S_i)$ is the weight for feature $f_k$ in story $S_i$. The weighting function used in our experiments is a *tf-idf* function specified by the following formula [2]:

$$f_k(S_i) = \frac{tf(f_k, S_i)}{tf(f_k, S_i) + 0.5 + (1.5 * \frac{len(S_i)}{asl})} \cdot \frac{\log \frac{n + 0.5}{sf_k}}{\log(n + 1.0)}$$

$tf(f_k, S_i)$ is the term frequency of feature $f_k$ in story $S_i$ , $asl$ is the average number of features in a story for the topic, $sf_k$ is the document frequency of the feature, $len(S_i)$ is the number of features in the story, and $n$ is the number of stories for the topic. In our experiments, $n$ and $sf_k$ are computed incrementally based on the stories already in the stream.

## 3.4 Language Models

A *language model* is a probability distribution that captures the statistical regularities of features (visual keywords or text words). Applied to novelty and redundancy detection, visual language modeling refers to the problem of estimating the likelihood that two news stories could have been generated by the same language model, given two language models of stories built on the keyframes. The similarity between two stories can be measured by the *Kullback-Leibler (KL) divergence* between two language models. It is based on the pairwise similarity between the current story and each previously broadcast story in the topic.

### *3.4.1 KL Divergence Distance Metric*

In the language model approach, a story is represented by a unigram feature distribution $\theta$. We assume that a news story is generated from a probabilistic model based on story $S_i$. In the language model, a multinomial model $p(f_k|\theta_i)$ over feature $f_k$ is estimated for each news story $S_i$ in the collection $C$.

A distributional similarity measure, *KL divergence* (or relative entropy), is used to measure the similarity between two stories.

$$R(S_i, S_j) = -KL(\theta_i, \theta_j) = -\sum_{f_k} p(f_k \mid \theta_i) \log(\frac{p(f_k \mid \theta_i)}{p(f_k \mid \theta_j)})$$

where $\theta_i$ is the language model for story $S_i$, which is a multinomial distribution. Here $p(f_k|\theta_i)$ is the probability of feature $f_k$ occurring in news story $S_i$, similarly for $p(v_k|\theta_j)$. The *KL divergence* can be regarded as a distance between distributions. The higher the similarity is, the more near-duplicate two keyframes are.

The simplest way to estimate $p(f_k|\theta_j)$ is the *Maximum Likelihood Estimation (MLE)*, simply given by relative counts:

$$p(f_k \mid \theta_i) = \frac{tf(f_k, S_i)}{\sum_{f_k} tf(f_k, S_i)}$$

where $tf(f_k, S_i)$ is the term frequency of feature $f_k$ in news story $S_i$. However, the problem of maximum likelihood estimation is that it will generate a zero probability if a feature never occurs in the story $S_i$, which will cause $KL(\theta_i, \theta_j)=\infty$.

Smoothing techniques are used to assign a non-zero probability of the unseen features and improve the accuracy of the feature probability estimation. Prior research on language model smoothing in text [2, 30, 33] showed that different smoothing methods highly affect performance. For language models, we mainly use *Bayesian* smoothing with *Dirichlet* priors and *Shrinkage*. Furthermore, we also implement a *Mixture Model*.

### *3.4.2 Dirichlet Smoothing*

This smoothing technique uses the conjugate prior for the multinomial distribution, which is the *Dirichlet* distribution. It automatically adjusts the amount of reliance on the features according to the total number of the features. For a Dirichlet distribution with parameters:

$$(\mu p(f_1 \mid C), \mu p(f_2 \mid C), \ldots, \mu p(f_n \mid C))$$

the posterior distribution using Bayesian analysis is:

$$p_\mu(f_k \mid \theta_i) = \frac{tf(f_k, S_i) + \mu p(f_k \mid C)}{\sum_{f_k} tf(f_k, S_i) + \mu}$$

$p(f_k|C)$ is the collection language model and $\mu$ is a parameter to adjust the degree of smoothing.

**Table 1. Topic information (#: number of stories)**

| English + Chinese | | | | | English | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **ID** | **Topics** | **#** | **CR** | **SR** | **ID** | **Topics** | **#** | **CR** | **SR** |
| 1 | APEC summit | 37 | 5 | 7 | 1 | Intelligence reform bill | 22 | 0 | 6 |
| 2 | Arafat health | 155 | 22 | 42 | 2 | Ebersol plane crash | 9 | 1 | 6 |
| 3 | Black Friday | 18 | 1 | 4 | 3 | Bush & Blair | 17 | 2 | 6 |
| 4 | Cabinet changes | 78 | 7 | 35 | 4 | Bush visited Canada | 10 | 1 | 4 |
| 5 | President Election | 184 | 12 | 20 | 5 | CIA in turmoil | 5 | 0 | 2 |
| 6 | War on Fallujah | 203 | 13 | 64 | 6 | NBA brawl | 32 | 0 | 21 |
| 7 | AIDS | 27 | 6 | 2 | 7 | Scott Peterson trial | 30 | 0 | 12 |
| 8 | Afghan hostage | 11 | 4 | 3 | 8 | Vioxx | 10 | 0 | 3 |
| 9 | Iraq problem | 158 | 12 | 21 | 9 | Vice president's health | 8 | 0 | 5 |
| 10 | Korean nukes problem | 21 | 2 | 5 | 10 | Veteran's day | 17 | 2 | 4 |
| 11 | Clinton library | 13 | 1 | 6 | **Chinese** | | | | |
| 12 | Iran nukes | 39 | 0 | 19 | 1 | Mine bombing | 14 | 4 | 6 |
| 13 | Mideast peace | 70 | 9 | 11 | 2 | China & Eastern Union | 24 | 6 | 3 |
| 14 | Bush second term plan | 31 | 2 | 6 | 3 | Hu Jintao visited South America | 35 | 10 | 5 |
| 15 | War on terrors | 27 | 1 | 12 | 4 | WTO | 11 | 2 | 2 |
| 16 | Thanksgiving | 22 | 2 | 1 | 5 | Falun Gong | 25 | 4 | 2 |
| 17 | Ukraine crisis | 68 | 9 | 20 | 6 | Yunnan air crash | 7 | 0 | 0 |
| | | | | | | **Total** | 1436 | 140 | 365 |

(CR: number of Completely Redundant stories SR: number of Somewhat Redundant stories)

### *3.4.3 Shrinkage Smoothing*

*Shrinkage smoothing* is a special case of the *Jelinek-Mercer* smoothing method, which involves a linear interpolation of the maximum likelihood model with the *n*-gram model [30]. Based on the assumption that a story is generated by sampling from three different language models: a story model, a topic model, and a model for the collection, the language model of a story is determined by:

$$p(f_k \mid \theta_s) = \lambda_s p(f_k \mid \theta_{ML_S}) + \lambda_T p(f_k \mid \theta_{ML_T}) + \lambda_C p(f_k \mid \theta_{ML_C})$$

using coefficients $\lambda_s$, $\lambda_T$, and $\lambda_C$ to control the influence of each model, where $\lambda_s + \lambda_T + \lambda_C = 1$. $\theta_{ML_T}$ and $\theta_{ML_C}$ are the maximum likelihood language models of the topic and collection respectively. In our experiments, the topic model is built on all presumed relevant stories for the topic and the collection model is built on all stories in the corpus.

### *3.4.4 Mixture Model*

A story is generated by the mixture of three different language models: a story-specific model $\theta_s$, a topic model $\theta_T$, and a model for the collection $\theta_C$. A *mixture model* [33] is based on the opposite assumption that features occurring more frequently in a story than in the background (topic or collection) should have higher probability in the story model. Therefore, the approach is to deduce the maximum likelihood story model, which is compared pairwise to each previously seen story model. Each feature in the story is generated by each of the three language models with probability $\lambda_s$, $\lambda_T$, and $\lambda_C$ respectively (where $\lambda_s + \lambda_T + \lambda_C = 1$).

$$p(f_k \mid \theta_{ML_x}) = \lambda_s p(f_k \mid \theta_s) + \lambda_T p(f_k \mid \theta_{ML_T}) + \lambda_C p(f_k \mid \theta_{ML_C})$$

To note, although equations of shrinkage smoothing and mixture model look similar, the model acquired and used to calculate *KL* divergence is different. Shrinkage smoothing increases the probability of features that occur frequently in the topic or in the collection if they occur less frequently in the story, while a mixture model decreases the probability of these features [33]. Similar to [2], the language model $\theta_s$ that maximizes the likelihood of the observed story, given fixed parameters, is computed using the technique described in [34].

## 3.5 Multimodality Fusion

In the proceed sections, we discuss different measures on individual feature to measure the similarity. However, for news stories, the reliance of either textual or visual features may not be sufficient since either feature can behave differently over time. Textual and visual concepts complement each other. The pure textual method may overlook the interactions between textural and visual information, i.e., where the visual contents determine the set of shots on which text summarization will be considered, but the textual information does not have a say about how the set of shots is selected. Robust and reasonable approaches should combine both textual and visual features to determine the degree of novelty in news stories while exploiting the significance of these features.

Motivated by this fact, the following measure which integrates the textual and visual features is proposed to detect the novelty/redundancy. We use a linear weighted fusion method for combining the similarity scores from speech transcripts and visual duplicates (i.e. *T* and *V*). Linear fusion model has been shown to be one of the most effective approaches to fuse textual and visual modalities in video retrieval (e.g. [31]). The similarity score is defined as:

$$R(S_i, S_j) = \alpha R_T(S_i, S_j) + (1-\alpha) R_V(S_i, S_j)$$

The measure of textual and visual features can be any method mentioned in previous sections. Their similarity scores are normalized in the range of [0, 1] for fusing. The weight $\alpha$ *is* used

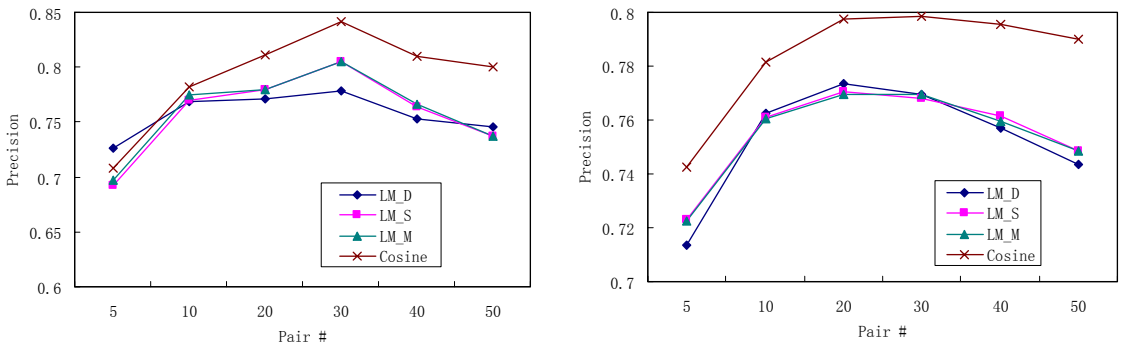



**Figure 3. Effects of near-duplicate keyframe detection with different thresholds**
**News stories are considered redundant if assessors marked it as (a) completely redundant (b) completely or somewhat redundant**

to control the influence of each feature. The linear weight among modalities is determined empirically. Although the adopted fusion method (linear fusion) is not novel, the purpose of this paper is not to discuss the fusion strategy, but to explore the feasibility of multimodal fusion to improve the performance for cross-lingual novelty and redundancy detection.

# 4. EXPERIMENTS

## 4.1 Dataset and Performance Metric

We selected all Chinese and English videos from the TRECVID-2005 cross-lingual news video corpus [26] as our data set which is around 127 hours. It includes 88 Chinese and 142 English news videos from five different sources (CCTV4, NTDTV, CNN, NBC and MSNBC). Arabic videos were ignored in our experiments because assessors were not familiar with Arabic and could not judge similarity. Due to the lack of standard story boundaries and topic annotation, we used the story boundaries from CMU Informedia [10] to segment video streams into stories, and the topics were manually labeled and annotated under the guidance of TDT (topic detection and tracking). The data set consisted of 1,436 news stories. To ensure the fairness of comparison, topics that contained less than four news stories were regarded as outliers and removed. There remained 33 topics. Among them, 10 topics had only stories reported by the English channels (ENG), while 6 topics were only broadcast over the Chinese channels (CHN). The detailed information is shown in Table 1.

The visual information is a set of representative keyframes extracted from video corpus. Note that each shot is represented by one keyframe. The shots with an anchor person were automatically detected and removed with CMU's detection method. The anchor person classifier was generically trained, using 3HVC color features, SVM with RBF kernel, all positive examples and 1/5 negative examples. The average accuracy on the held-out set of TRECVID 2005 development data is 90.5%. The keyframes having split screen with anchor and another person (e.g. interviewee) are not treated as anchorperson. Totally, there are 19,621 representative keyframes.

Different people have different definitions of redundancy and different redundancy thresholds. In order to alleviate such inconsistency, similar to [33], we classified news stories into three classes in this paper: *completely redundant*, *somewhat redundant*, and *novel*. If a news story contained no new information, this news story is regarded as *completely redundant*. A news story which is a review or a duplicate news story with very minor adjustment of previously delivered news story is completely redundant news story. Only when both visual and textual contents of a news story are totally covered by previous stories, it is treated as completely redundant. Those news stories that have some new information and contain many redundant contents are marked as *somewhat redundant* news stories. Somewhat redundant stories usually convey the gradual development of a theme. News stories where most of the contents are new are marked as *novel* stories which indicate the emergence of new themes.

To analyze the performance of the novelty/redundancy detection scheme, two undergraduate students from CMU as the assessors were asked to watch stories and judge one topic at a time. News stories in each topic were ordered chronologically. The assessors were requested to label the news stories with a judgment (completely redundant, somewhat redundant or novel). The information of completely redundant and somewhat redundant stories is listed in Table 1.

To factor out the effect of the redundancy threshold, we evaluate the effectiveness of redundancy detection by comparing the average precision and recall figures. A redundancy score is calculated for each news story by comparing it with preciously delivered stories. News stories are ranked by their redundancy scores. *Precision* and *recall* are commonly used metrics in information retrieval. Let $\boldsymbol{G}$ be the ground truth set of redundant stories and $\boldsymbol{D}$ be the detected one.

$$R\_Recall = |G \cap D| / |G| \qquad R\_Precision = |G \cap D| / |D|$$

Moreover, we also calculated the average precision over all recall levels.

## 4.2 Effect of Near-Duplicate Keyframe Detection

To detect near-duplicate keyframes, the local interest points of each keyframe were extracted by Lowe's DoG detector [15] and described by PCA-SIFT [12], which is a 36 dimensional vector for each interest point. One-to-one symmetric matching [20] based on local interest points was used to detect whether two keyframes are near-duplicates. There is no ground truth for the NDK detection for TRECVID-2005 dataset. In our previous experiments on a subset of TRECVID-2004, the accuracy is around 80-85%. The number of matching pairs among interest points is the major parameter. In our experiments, two keyframes were treated as a NDK pair if the matching pairs of local interest

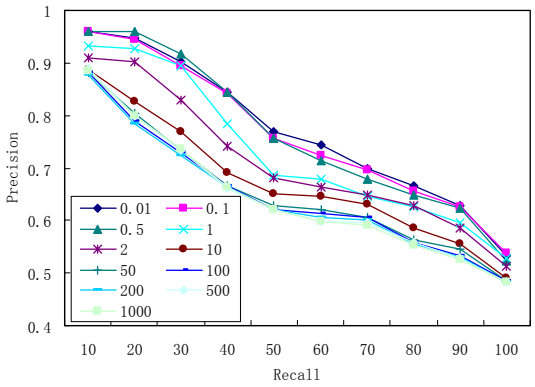
**Table 2. Visual Keywords Information with Different Settings (Two keyframes are regarded as NDK if the number of matching pairs of their local interest points is great than # pairs)**

| # Pairs | Non-NDK | NDK | NDK_Group |
|---|---|---|---|
| 5 | 10,223 | 9,398 | 2,007 |
| 10 | 11,442 | 8,179 | 2,261 |
| 20 | 12,420 | 7,201 | 2,263 |
| 30 | 13,229 | 6,392 | 2,139 |
| 40 | 13,978 | 5,643 | 1,987 |
| 50 | 14,564 | 5,057 | 1,837 |

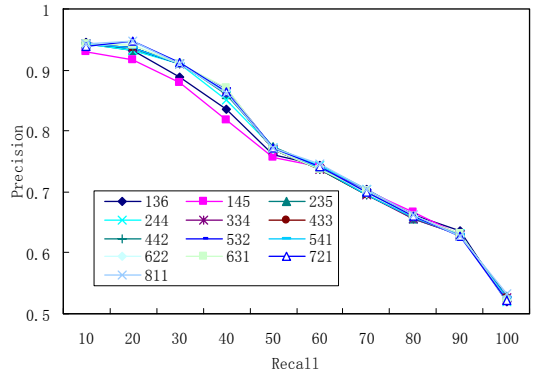
points between two keyframes are above a certain threshold. The whole near-duplicate keyframe list was generated by transitive closure based on the information from every two keyframes, which forms a set of NDK groups. The keyframes in each NDK group are very similar, that is, they represent one visual concept, and are regarded as one visual keyword. The information of visual keywords under different settings is listed in Table 2. For example, when the threshold (the number of matching pairs of interest points) is set to 20, there are 7,201 NDK in the data set, which form 2,263 groups, and 12,420 Non-NDK. So the vocabulary size is 14,683 (12,420 + 2,263).

As the near-duplicate keyframes are automatically detected, instead of manually labeled, different settings of threshold will affect the number of visual keywords and finally affect the performance of the novelty detection on keyframes. We compare the performance of Cosine distance on visual features (Cosine) and visual language models with Dirichlet smoothing (LM_D), with Shrinkage smoothing (LM_S) and Mixture Model (LM_M). The average precision over different recall levels with different thresholds (the numbers of matching pairs among interest points) of near-duplicate keyframes is shown in Figure 3.

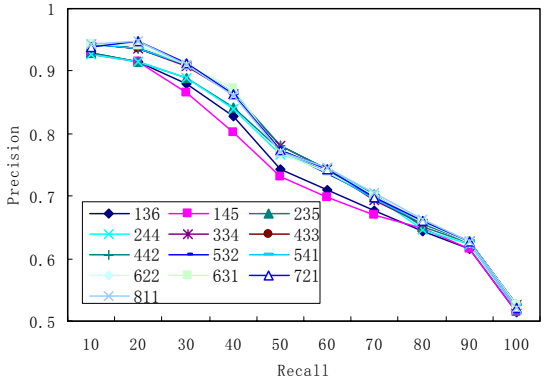
The performance is dependent on the accuracy of near-duplicate keyframe detection. In general, the performance increases at first and then drops a lot as the threshold setting of matching pairs between two keyframes increases. When the number of matching pairs is small (e.g. 5), more keyframes will be regarded as NDK. There is the case that some NDK should belong to different NDK groups are grouped together by transitivity and treated as one NDK group. And even Non-NDK are included into the groups and regarded as NDK due to the low threshold. It can be seen from Table 2 that the number of detected NDK is large. And the result is a little noisy which affects the distribution of visual keywords. Therefore, it is not strange that the overall performance is poor. On the contrary, as the threshold is high, for instance, two keyframes can be regarded as NDK only when they have at least 50 matching pairs of interest points, the number of NDK drops. Some NDK belonging to one group are separated into a couple of groups, so they will be treated as different visual keywords. And many NDK are falsely detected as Non-NDK due to the high threshold, and will not be included into the NDK groups. So the distribution of visual keywords deviated from the original one. The performance drops when the setting is too high. But when the number of matching pairs is around 30, NDK can be correctly detected and grouped, which is approximately the true distribution of visual keywords. It achieves the best performance. In the later experiments, we used 30 as the default setting.




**(a) Dirichlet smoothing with different settings for parameter $\mu$**




**(b) Shrinkage smoothing with different settings for parameter ($\lambda_S$, $\lambda_T$, $\lambda_C$). (532 represents that the weights of $\lambda_S$, $\lambda_T$, $\lambda_C$ are 0.5, 0.3 and 0.2 respectively)**

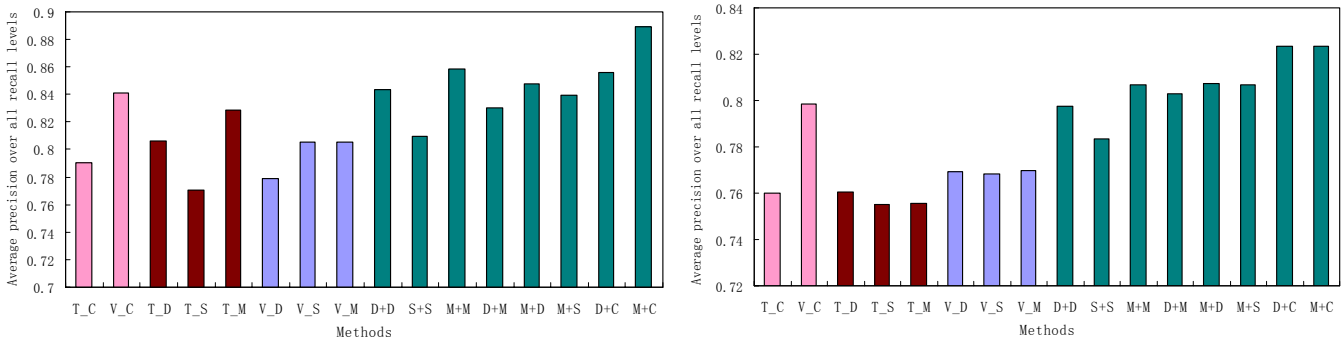



**(c) Mixture model with different settings for parameter ($\lambda_S$, $\lambda_T$, $\lambda_C$). (532 represents that the weights of $\lambda_S$, $\lambda_T$, $\lambda_C$ are 0.5, 0.3 and 0.2 respectively)**

**Figure 4. Performance of smoothing for language models build on visual features.**

## 4.3 Effect of Smoothing

Traditional language models on text are sensitive to the smoothing methods and the parameter settings [30]. However, visual language models have their special properties, such as small number of visual keywords in each story and large number of Non-NDK. To study the sensitivity of visual language models, we compare the visual language models with different smoothing parameter settings. Figure 4 shows the precision-recall graphs of




**(a) completely redundant** **(b) completely or somewhat redundant**

**Figure 5. Performance comparison of different uni-modal and multi-modal approaches with different colors (Cosine distance, Language models on text, Language models on visual features, Fusion of text and visual features)**

**The first eight methods are unimodal methods on textural (T) and visual (V) features respectively, while the last eight methods are the fusion of textual and visual features represented by $Measure_T$+$Measure_V$ pairs, in which measures are denoted as C – Cosine, D – Dirichlet smoothing, S – Shrinkage smoothing, M – Mixture Model.**

smoothing. For Shrinkage smoothing and Mixture Model, the weights of story, topic and corpus are represented by the format: $\lambda_s\lambda_T\lambda_C$. Their summation is equal to 1. For example, 622 represents that the weights of $\lambda_s$, $\lambda_T$, $\lambda_C$ are 0.6, 0.2 and 0.2 respectively. Generally speaking, Dirichlet smoothing is sensitive to the parameter $\mu$, while Shrinkage smoothing is relatively stable with the weights of story, topic and corpus models ($\lambda_s$, $\lambda_T$, and $\lambda_C$).

For Dirichlet smoothing (Figure 4(a)), the smoothing performance is story-dependent. The coefficient $\alpha_d$ controlling the probability mass assigned to unseen visual keywords is closely related to the visual length of story: $\alpha_d = \mu/(\Sigma tf(f_k,S_i)+\mu)$. The relative weighting of visual keywords is emphasized when the parameter $\mu$ is small. It achieves good performance when $\mu$ is small. As $\mu$ becomes large, the coefficient $\alpha_d$ tends to 1. The weighting of unseen visual keywords has little impact. Furthermore, the number of visual keywords in one story is relatively small compared to traditional textual words, so the situation is more conspicuous. Therefore the performance is poor when the parameter $\mu$ is large.

For Shrinkage smoothing and Mixture Model (Figure 4 (b) and (c)), the parameters ($\lambda_s$, $\lambda_T$, $\lambda_C$) are same for all stories, which is story independent. When $\lambda_s$ is small and $\lambda_C$ is high, visual keywords term weights become insignificant. The probability of visual keywords is mainly determined by the topic and background corpus visual models, which cannot provide an accurate estimation. So the performance is not good. On the contrary, when $\lambda_s$ is high and $\lambda_C$ is small, it emphasizes more on the relative visual term weighing. The probability of visual keywords is controlled more by the story visual model, and less by the corpus visual model. Therefore, the performance improves. And Mixture Model is more sensitive to the parameter setting than Shrinkage smoothing.

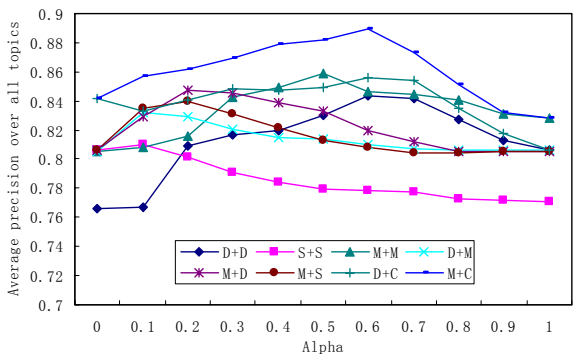
## 4.4 Performance Comparison for Measures

To compare the performance of different approaches, we compare Cosine distance and language models on textual and visual features respectively. The visual features are the visual keywords (NDK and Non-NDK) while the textual features are a list of words extracted from speech transcripts by an automatic speech recognition system (ASR) at LIMSI [7]. There are 42 videos (out of 230) having no ASR or corresponding machine translation. To eliminate the effect of absence of text features, we supplemented them with speech recognition software [10] and translated them with Google translation [8] so that all stories have English transcripts. After data preprocessing such as word stemming and stop-word removal, there are 12,428 unique words.

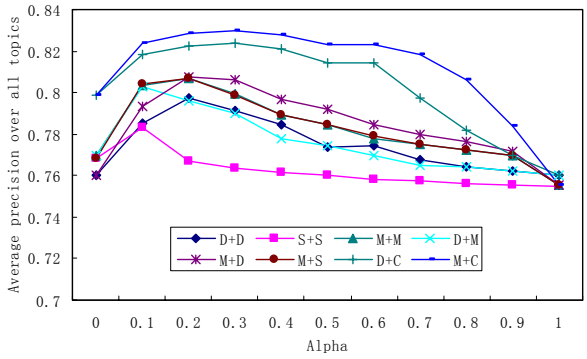
Figure 5 summarizes the comparison of different approaches, in which the Cosine distance is labeled in pink, while language models built on textual and visual features are displayed in brown and blue, respectively. We choosed the Cosine distance on text as the baseline because it has been a standard, robust measure in many tasks. In general, Cosine distance on keyframes outperforms the other approaches. Visual language models demonstrate promising performance. Due to the small number of keyframes in each story and the large number of Non-NDK, visual language models have dissimilar results compared to text language models.

For completely redundant stories (Figure 5(a)), text-based methods can still identify many completely redundant stories. Different smoothing techniques play an important role in the probability estimation for text words. Text-based Dirichlet smoothing and Mixture Model show better performance than the text-based Cosine distance. Even if speech recognition and machine translation might falsely detect and translate speech transcripts into incorrect words, a lot of completely redundant stories can still be correctly matched by text. Visual information (NDK information) is a strong indication for completely redundant stories, especially for the cross-lingual broadcast domain. Although the special properties of keyframes make the probability estimation less effective, visual language models show their potential.

When both completely redundant and somewhat redundant stories are treated as redundant stories (Figure 5(b)), visual language models outperform text-based measures because visual information provides important and direct cues to measure similarity among somewhat redundant stories. The difference among smoothing techniques is reduced.




**(a) completely redundant**




**(b) completely or somewhat redundant**

**Figure 6. Effect of fusion factor $\alpha$**

## 4.5 Performance Comparison for Fusion

Because news videos convey information both in the audio and visual track, similarity measures on either textual or visual features may not be sufficient. A robust solution takes into account both textual and visual features while exploiting the significance of these features. We compare different combinations to study the possible improvement from fusing texts and keyframes together. The measures are represented by $Measure_T$+$Measure_V$ pairs (green bars in Figure 5). For example, M+D denotes that this method uses a Mixture Model to measure the text and Dirichlet smoothing to compare visual features, and then combines the results. For each combination, we implement different fusion factors and report only the best result.

Figure 5 illustrates that fusion of textual and visual features improves performance substantially. Generally the measures fusing textual and visual information achieve better performance than the individual measures before fusion. The best performance substantially outperforms the Cosine distance on keyframes. Textual features and visual features complement each other. Especially for stories with similar textual concepts but different visual information, or with similar visual appearance but inconsistent text contents, fusing multiple modalities can be beneficial.

The effect of the fusion factor $\alpha$ is shown in Figure 6. The general trend is the performance increases at first and then drops. When $\alpha$ is zero, textual features have no effect and the measures rely only on visual information. When the fusion factor increases, the textual similarity takes part. Generally we can see a performance improvement as the factor increases, which proves that the textual features provide a meaningful complement for the visual contents. For two stories having high visual information, if their textual contents are also similar, it reinforces the confidence of redundancy. On the contrary, due to lack of information, although two stories have the similar visual contents, they may discuss different subtopics under the same topic. Under this situation, fusion of textual and visual features potentially low the score of redundancy, which provides a more reasonable evaluation. As the fusion factor increases, the textual similarity begins to dominate and the performance drops. Obviously having high textual similarity alone is not a strong indication of redundancy. Two stories may discuss the same event, but they may carry totally new visual contents. In this case, they can not be treated as completely redundant. We can arrive at the conclusion only if they also have high visual similarity. Overall, the combination of textual and visual similarity gives a more accurate evaluation for novelty and redundancy.

# 5. CONCLUSION

Novelty and redundancy detection for cross-lingual broadcast news stories is a substantial task for news story search, retrieval, and tracking. Differing from traditional text-based news articles, news videos include both audio and visual channels. Visual information provides useful cues for the similarity measure, and complements with the noisy textual contents in the cross-lingual scenario. In this paper, we explore the feasibility of building visual language models at the keyframe level, and compare different approaches using textual and visual features for the novelty and redundancy detection of cross-lingual news stories. Furthermore, we also fuse the textual and visual features to boost the performance. Experiments on cross-lingual news video corpus TRECVID-2005 show that:

- Due to the special properties of keyframes (the small of keyframes in each story and the large number of Non-NDK), approaches on visual information demonstrate different performance compared with traditional methods on text.
- Cosine distance on visual features performs better than other measures on unimodality (text or visual features).
- Language models of visual features are effective, but depend on the accuracy of near-duplicate keyframe detection.
- Visual language models are less sensitive to smoothing techniques, and have better performance than text-based measures when both completely and somewhat redundant stories are counted in the redundancy metric.
- A combination of textual and visual information improves the performance, which achieves better performance than individual ones.

Our research can be easily applied to other applications, such as web video similarity measure, video tracking and threading. Scalability is another important issue to further study. Our method will be tested in the TRECVID-2006 and other data sets to verify its scalability. In the future we will investigate new language models specially tailored to the unique characteristics of visual features. Currently, we use Google translation as a supplement. We will explore combing different machine translation tools such

as Google translation and SYSTRAN machine translation to boost the quality of noisy transcripts, and eventually improve the performance of novelty detection. Furthermore, web resources provide abundant information, which will beneficial for translating named entities and out-of-vocabulary (OOV) words, and mapping broadcast news videos and news articles in the web. They will contribute to the performance of novelty detection for cross-lingual news videos.

## 6. ACKNOWLEDGEMENT

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 118905) and a grant from City University of Hong Kong (Project No. 7002112).

We'd like to thank Wan-Lei Zhao for the NDK detection.

## 7. REFERENCES

[1] J. Allan, (ed.) Topic Detection and Tracking: Event-based Information Organization. *Kluwer Academic Publishers, 200*2.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and Novelty Detection at the Sentence Level. *ACM SIGIR'03*.

[3] T. Brants, F. Chen, and A. Farahat. A System for New Event Detection. *SIGIR'03*, Canada, Jul. 2003

[4] S-F. Chang and et al. Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction. *TRECVID 2005*, Washington DC, 2005.

[5] S. C. Cheung and A. Zakhor. Efficient Video Similarity Measurement with Video Signature. *IEEE TCSVT*, 2003.

[6] E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing Personalized Newsfeeds via Analysis of Information Novelty. *WWW'04*, USA, 2004, pp. 482-490.

[7] J. L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 2002.

[8] Google translation. http://translate.google.com.

[9] W. H. Hsu and S-F. Chang. Topic Tracking across Broadcast News Videos with Visual Duplicates and Semantic Concepts. *ICIP'06*, Atlanta, GA, October 2006.

[10] Informadia. Available: http://www.informedia.cs.cmu.edu.

[11] A. K. Jain, A. Vailaya, and W. Xiong. Query by Video Clip. *ACM Multimedia Syst. J.*, vol. 7, pp. 369–384, 1999.

[12] Y. Ke, R. Sukthankar, and L. Huston. Efficient Near-Duplicate Detection and Sub-Image Retrieval. *ACM MM'04*.

[13] L. S. Larkey, F. Feng, M. Connell, and V. Lavrenko. Language-specific Models in Multilingual Topic Tracking. *SIGIR'04*, UK, Jul. 2004.

[14] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-Lingual Relevance Models. *SIGIR'02*, pp. 175-182, 2002.

[15] D. Lowe. Distinctive Image Features from Scale-Invariant Key Points. *IJCV*, vol. 60, pp. 91-110, 2004.

[16] K. McDonald and A. F. Smeaton. A Comparison of Score, Rank and Probability-Based Fusion Methods for Video Shot Retrieval. *CIVR'05*, pp. 61-70.

[17] D. Metzler, Y. Bernstein, W. Croft and et al. Similarity Measures for Tracking Information Flow. *CIKM'05*.

[18] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann and J. Curtis. Large-Scale Concept Ontology for Multimedia. *IEEE Multimedia Magazine*, 13(3), 2006.

[19] A. Natsav, M. R. Naphade, and J. Tesic. Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples. *ACM MM'05*, pp. 598-607.

[20] C-W. Ngo, W-L. Zhao, Y-G. Jiang. Fast Tracking of Near-Duplicate Keyframes in Broadcast Domain with Transitivity Propagation. *ACM MM'06*, USA, Oct. 2006.

[21] Y. Peng and C-W. Ngo. Clip-based Similarity Measure for Query-Dependent Clip Retrieval and Video Summarization. *IEEE Trans. on CSVT*, vol. 16, no. 5, May 2006.

[22] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. *ACM SIGIR'98*.

[23] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. Smeulders. The Challenge Problem of Automated Detection of 101 Semantic Concepts in Multimedia. *ACM MM'06*, pp. 421-430.

[24] T. Westerveld. Using generative probabilistic models for multimedia retrieval. *Ph. D thesis*, 2004.

[25] T. Westerveld and A. P. Vries. Multimedia Retrieval using Multiple Examples. *CIVR'04*.

[26] TRECVID 2005. http://www-nlpir.nist.gov/projects/trecvid/.

[27] Wikipedia. http://en.wikipedia.org/wiki/Youtube.

[28] X. Wu, C-W. Ngo, and Q. Li. Threading and Autodocumenting News Videos. *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 59-68, March 2006.

[29] Y. Yang, J. Zhang, J. Carbonell and C. Jin. Topic-conditioned Novelty Detection. *ACM SIGKDD'02*, Canada.

[30] C. Zhai and J. Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. *ACM SIGIR'01*, USA, pp. 334-342, Sep. 2001.

[31] Y. Zhai and M. Shah. Tracking News Stories across Different Sources. *ACM MM'05*, Singapore, Nov. 2005.

[32] D-Q. Zhang and S-F. Chang. Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. *ACM MM'04*, USA, Oct. 2004.

[33] Y. Zhang, J. Callan, and T. Minka. Novelty and Redundancy Detection in Adaptive Filtering. *ACM SIGIR'02*, 2002.

[34] Y. Zhang, W. Xu, and J. Callan. Exact Maximum Likelihood Estimation for Word Mixtures. *Text Learning Workshop at the Int. Conf. on Machine Learning (ICML)*, 2002.

[35] X. Zhu, J. Fan, A. K. Elmagarmid, and X. Wu. Hierarchical Video Content Description and Summarization Using Unified Semantic and Visual Similarity. *Multimedia System*, vol. 9, no. 1, 2003, pp. 31-53.