AT&T Research at TRECVID 2010

Zhu Liu, Eric Zavesky, Neela Sawant*, Behzad Shahraray

AT&T Labs Research, 200 Laurel Avenue, Middeltown, NJ 07748
*College of Information Sciences and Technology, Penn State University, University Park, PA 16802
{zliu,ezavesky,behzad}@research.att.com, *nks125@ist.psu.edu

ABSTRACT

AT&T participated in two tasks at TRECVID 2010: content-based copy detection (CCD) and instance-based search (INS). The CCD system developed for TRECVID 2009 was enhanced for efficiency and scale and was augmented by audio features [1]. As a pilot task, participation in INS was meant to evaluate a number of algorithms traditionally used for search in a fully automated setting. In this paper, we report the enhancement of our CCD system and propose a system for INS that attempts to leverage retrieval techniques from different audio, video, and textual cues.

1. INTRODUCTION

TRECVID started as a video track of TREC (Text Retrieval Conference) in 2001 to encourage research in automatic segmentation, indexing, and content-based retrieval of digital video and in 2003 it became an independent evaluation [2]. TRECVID 2010 presented a forum for evaluating traditional tasks like content-based copy detection (CCD), high-level concept classification or semantic indexing (SIN), and event detection (SED) and introducing instance-based search (INS), known-item search (KIS), and multimedia event detection (MED) as new tasks. Following traditions, the TRECVID community pushed the envelope of each task by introducing a highly heterogeneous dataset captured from the Internet Archive into a dataset referred to as IACC.1. In this paper, we describe our work for the CCD and INS tasks and briefly discuss initial reactions from the formal TRECVID evaluations.

Instance-based search is pilot task in TRECVID 2010 that reuses part of the TRECVID 2009 dataset from the Sound & Vision Dutch archive (S&V). Most of the video in this archive can be described as documentary and educational. The major challenge that distinguishes the INS task from traditional search is the definition of a query with a visual object that is explicitly outlined in several images that closely correspond to each other (i.e. the same instance). While some textual information is provided for this query image, the focus (and intent of the task) is to find similar instances of that query object with only a basic description. Traditionally, tasks focusing on object detection and retrieval have used datasets that focus on a single object with many similar appearances (like correctly classifying a coffee cup) [3]. However, in recent years, these scenes containing these objects have become quite realistic although they still focus on still-frame recognition [4]. As a pilot task similar to this latter evaluation, the INS task was introduced in 2010 to measure retrieval capabilities for the S&V videos.

This paper is organized as follows. Section 2 gives a detailed description of the content-based copy detection system. Section 3 addresses our work for fully automated instance based search. Evaluation results from TRECVID 2010 are presented and discussed in section 4, and we summarize our conclusions in Section 5.

2. CONTENT-BASED COPY DETECTION

name	description
nofa.1	SIFT+audio LSH; fused 1:25 (0.5 threshold)
balanced.2	SIFT+audio LSH; fused 1:25 (0.2 threshold)
nofa.3	SIFT+audio LSH; fused 2:3 (0.5 threshold)
balanced.4	SIFT+audio LSH; fused 2:3 (0.2 threshold)

Table 1: CCD run names and descriptions.

2.1 Task Overview

Video copy detection is essential for many applications, for example, discovering copyright infringement of multimedia content, monitoring commercial air time, querying video by example, etc. The goal of video copy detection is to locate segments within a query video that are copied or modified from an archive of reference videos. Usually the copied segments are subject to various audio/visual transformations, which make the detection task more challenging. TRECVID 2010 CCD considers the following 8 categories of visual transformation and 7 categories of audio transformations.

- TV1: Simulated camcording
- TV2. Picture in Picture (PiP)
- TV3. Insertions of pattern
- TV4. Strong re-encoding
- TV5. Change of gamma
- TV6. Decrease in quality: a mixture of 3 transformations among blur, gamma, frame dropping, contrast, compression, ratio, white noise.
- TV8. Post production: a mixture of 3 transformations among crop, shift, contrast, text insertion, vertical mirroring, insertion of pattern, picture in picture.
- TV10. Combinations of 3 transformations chosen from T1 - T8.
- TA1: No audio transformation (nothing)
- TA2: MP3 compression
- TA3: MP3 compression and multiband companding
- TA4: Bandwidth limit and single-band companding
- TA5: Mix with speech
- TA6: Mix with speech, then multiband companding

- TA7: Bandpass filter, mix with speech, and compression

This year, TREVID only evaluated audio+video queries, which consist of the aligned versions of transformed audio and video queries. Each original query is expanded to 8x7=56 versions of audio+video queries using the 56 different combinations of audio and video transformations.

Starting from a strong performance from our video-only system in TRECVID2009 [1], we attempted to reduce algorithm runtime and add transformation robustness with alternative feature representations. An audio hashing component, based on prior work in TRECVID [5], was also added to our system to gracefully recover from queries where the video content was too severely degraded. While these two approaches are quite robust on their own, one challenge was efficiently resolving the sample rate and score ranges of each approach. In total, we submitted 4 runs for CCD evaluation, which are listed in Table 1.

2.2 CCD System

Figure 1 illustrates a high level overview of our audio/visual CCD system. We developed the visual-based and audio-based approaches independently, and each module produces a CCD run. The fusion step is a straightforward linear weighting mechanism to combine the audio and video runs together. By adjusting weights, we can make the overall CCD run be more influenced by either the audio or the visual modality.

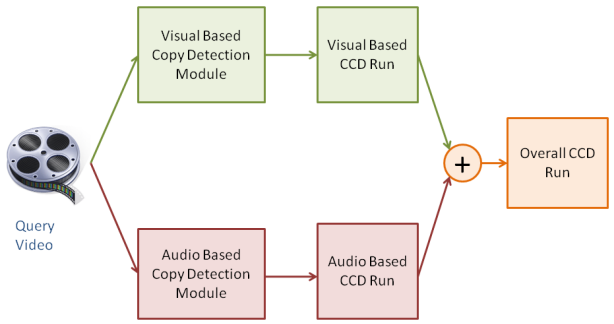

Figure 1: Diagram of the audio/visual CCD system.

2.2.1 *Video subsystem*

The video based CCD processing consists of two parts as indicated by different colors. Figure 1 shows the system overview. The top portion illustrates the processing components for the query video, and the bottom portion shows the processing stages for the reference videos.

For details of the visual processing, please consult our notebook paper from TRECVID 2009 [1]. In our 2009 CCD system, LSH indexing is sorts all LSH hash values of the reference videos, and the LSH query is a binary search, whose complexity is in the order of log(N), where N is the cardinality of reference SIFT points. This year, we focused on improving the scalability of LSH indexing and query evaluation, indicated by the blocks with bold font in Figure 2. We improved the efficiency of LSH indexing and query evaluation by a standard hashing approach (see Figure 3).

We used a classic hash table to index computed LSH values for gains in speed and reduced resource usage. In our system, 24 hash values are extracted for each SIFT feature,

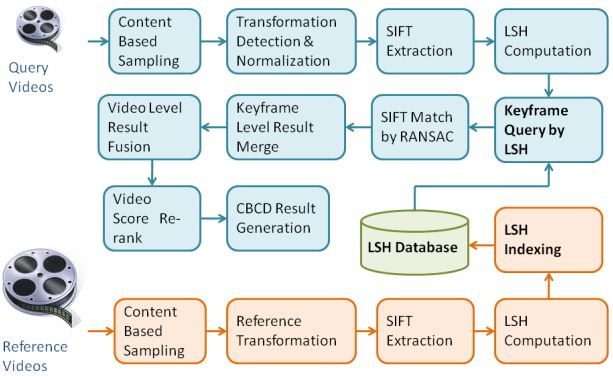

Figure 2: Overview of the proposed visual CCD algorithm .

where each hash value contains 32 bits. For two SIFT feature points, as long as any of the corresponding 24 pairs of LSH values match, these two points are claimed a matched pair. While indexing, we use 24 hash tables to index the 24 sets of LSH values. The size of hash table depends on the number of unique LSH values, and the tolerance of hash value conflicts. In this system, we set the size of hash table, M, to 16 millions. The original LSH value is mapped to the entry in the table by a hashing function (32 bit integer to [0, 16M) mapping), and conflicting entries are linked through pointers (e.g., Entry 1′ and Entry 1″). The detailed data structure of each entry is shown on the right hand side of Figure 3. The first field keeps the original LSH value, the second field counts the number of reference SIFT points that are mapped to this entry, the third field saves the list of these reference SIFT IDs, and the last field is a pointer to the next entry, in case there is a conflict. Normally, the last field is set to NULL.

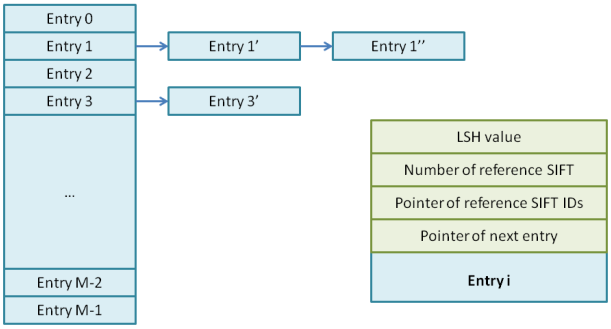

Figure 3: LSH indexing.

While indexing all SIFT LSH values in the reference dataset, the 24 hash tables are populated, and the arrays of reference SIFT IDs in each entry is sorted based on their {video, frame, SIFT} point IDs. Compared to the binary search method, the new implementation maintains a near constant time query complexity, and it increases the LSH query speed significantly.

Once we have the hash tables, we can trim them based on the number of reference SIFT points. If this number is too high, it means the corresponding SIFT feature is not descriptive, and it can be removed from the table. The trimming process benefits the overall system in two ways: 1) reduce the query speed and 2) improve the robustness of

SIFT based query.

2.2.2 *Audio subsystem*

The audio based CCD system was inspired by the approach reported by CRIM in TRECVID 2009 [5]. We use the energy difference between the sub-bands as the hash value for each frame. In this work, the original audio signal is re-sampled in 8KHz. Each frame is 32 milliseconds long, and adjacent frames overlap by 22 milliseconds (which leads to 100 frames per second). We considered 17 subbands, and 16 bit hash value. The 17 subbands are in mel-scale between 133 Hz and 3000Hz.

The same hash indexing method described in the last section is adopted in the audio approach. Instead of having 24 tables in visual approach, we only need one table, and the size of the table is fixed to 32K entries (due to 16 bit hash values).

The main challenge is how to efficiently implement the audio hash query. Given that there are millions of audio frames, and only 32K different hash values, a large number of hits are frequent for many query hash values. Another difference from a SIFT LSH query is that an audio query needs to consider temporal information. The basic approach uses a sliding window and counts how many audio frames match within this window. A direct matching implementation is too slow, and in our system, we adopted a much faster approach, again based on the hashing technique.

Figure 4 shows the detail of this query method. An audio query is shown at the top left corner, assuming there are N frames in this query. Hits for each query frame can be easily retrieved from the LSH indexing method shown in Figure 3. The results are plotted in the hit matrix under the query audio. Each colored block represents a hit, which is specified by a reference video ID and a reference audio frame number. Then for each hit, we compute the difference of query audio frame and reference audio frame (so called Q-R delta). We use the combination of reference video ID and the Q-R delta value as key to create a new counting hash table, where the hash value represents the frequency of hits for a certain reference video with a certain temporal offset. Using this technique, we only need to scan the hit matrix once to achieve the same query result as sliding window.

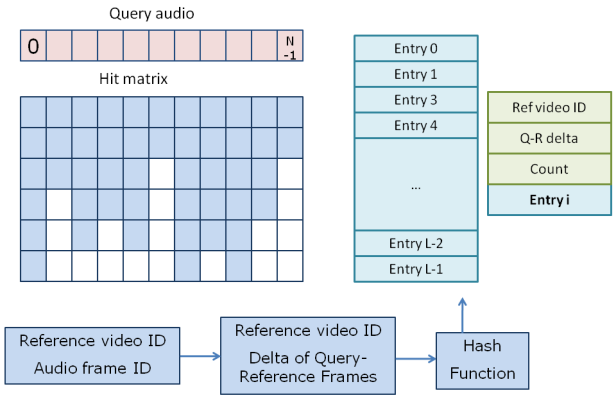

Figure 4: Audio Hash value query.

2.3 Fusion of audio and video information

We fuse the audio and visual-based CCD results at the final stage. For each query, audio-based and visual based CCD modules each reports a list of hits. Each hit specifies a reference segment and a score. When the two lists are merged, overlapping reference segments are grouped together, and its corresponding score is set to a weighted sum of the audio-based score and the visual-based score. The weight of audio-based hit score is set to 1.0, and the weight of visual-based hit score is adjustable, noted by ω.

While generating runs, we only report the best hit whose scores are higher than a certain threshold τ. For no false alarm profile, we also require that the score of the best hit is significantly higher than that of the second best hit (at least 1.5 times).

3. INSTANCE-BASED SEARCH

name	description
edge.p4	image CCD
3g.p3	CCD, text, dominant concepts
hspa.p2	CCD, text, dominant concepts and then face prioritization
lte.p4	RANSAC of CCD, text, dominant concepts and then face prioritization

Table 2: INS run names and descriptions.

3.1 Task Overview

Attempting to evaluate the benefits of our content-based copy detection system in a new forum, we also participated in the instance-based search task. As formulated, we interpreted the instance search task as a hybrid between object retrieval and traditional multimedia search, as illustrated below in Figure 5. Specifically, with a query being defined with a textual query category, a brief textual excerpt, and an image region, there were only two core modalities that could be leveraged: visual and textual. The following sections describe the different search methods employed for these two modalities, the attempted fusion strategies, and finally evaluation results.

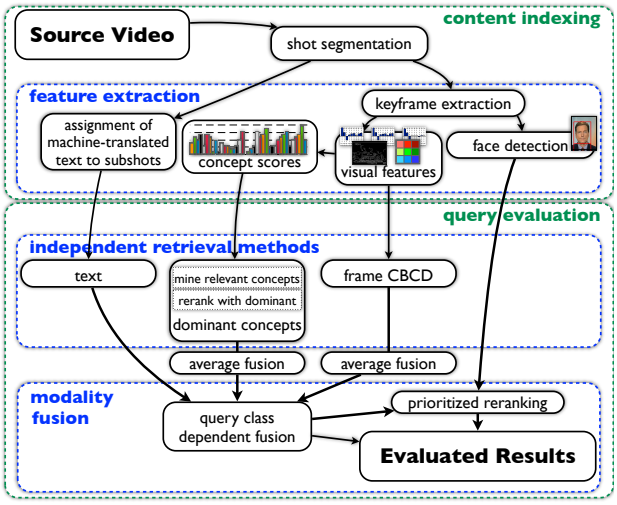

Figure 5: Overview of INS system from indexing (feature extraction) to query evaluation (search with multiple modalities and result fusion).

3.2 Search Methods

Multimedia retrieval systems work best when they take advantage of many different modalities during query evaluation. In the INS task, only a text phrase and a region-defined image are available, but these cues can be analyzed in multiple ways. It should be noted that although final evaluation is performed with the set of master shot boundaries provided by NIST, each search operation was actually executed using smaller subshots segments (and their representative keyframes) extracted by our own video segmentation algorithm [6]. Figure 5 demonstrates the contribution of multiple search methods and subsequent fusion across multiple modalities, as discussed in section 3.3.

3.2.1 Text

Our implementation of textual queries involves two straightforward steps: part of speech (POS) tagging and query phrase evaluation. After POS tagging [7], we use a points-based system to select different phrases for query formulation: noun phrases (i.e. proper nouns), nouns, and finally verbs. Although tagged, we do not process exclusionary modifiers (i.e. not, only, without), which was acceptable for this year's INS queries. Each textual query is evaluated with Lucene, which uses a combination of a vector space model and a binary model for result scoring. Textual documents are constructed form the machine-translated text that overlapped any part of each subshot. After a textual search, ordered results are saved for subsequent fusion or verification steps.

3.2.2 Content-based copy detection

Image queries in the INS task are actually defined by regions of one or more frames illustrated in Figure 6. The labeling of a region of interest allows the query to identify both traditional objects, like people and vehicles, as well as those with non-contiguous appearances, like the "zebra stripes" in an urban crosswalk. Fortunately, the content-based copy detection system described earlier in section 2.2 is capable of accepting arbitrary region definitions. Only two modifications were made for evaluating INS queries. The first is to prune interest points outside of the defined regions after detection and before query evaluation. This additional step allows the interest point detector to still operate on a multi-scalar nature but it should reduce noisy interest points that are not part of the primary object. Informal experiments demonstrated that including all interest points found with background regions rendered as various solid colors created too many false alarms along edges and approaches did not perform as well. The second modification is the separation of RANSAC verification stage into a secondary process. By isolating the verification of the visual content of two images, we are able to use this process to rerank results from other searches methods discussed in this section.

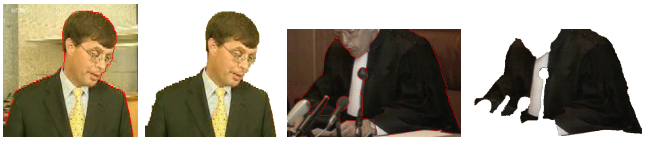

Figure 6: Examples of objects definitions and their regions for the INS task (topics 9003 and 9015).

3.2.3 Dominant semantic concepts

Semantic concepts are a popular way to represent the semantic bridging between low-level, machine-readable image features and high-level textual descriptions of content. Although modern efforts now seek to capture complex logical relationships between visual semantic concepts, the concepts and labels from two early efforts in this field (LSCOM [8] and MediaMill101 [9]) were used to train a lexicon of 474 semantic concepts on the TRECVID 2005 development corpus. Each concept score is a probabilistically normalized average of low-level classifiers using grid color moments, Gabor textures [10], edge direction histograms [11], and quantized SURF visual words [12]. Using these concept classifiers, all of the S&V test set and each query image was evaluated for each classifier. Scores below a certain threshold (here 0.05) were immediately pruned and the other scores are stored as representations of each subshot's keyframe.

Contrary to traditional retrieval systems that attempt to map the query text into a concept lexicon, we use the query image itself as a method for discovering the dominant (i.e. salient) visual concepts within an image. This approach was warranted because of the short textual topic definitions and large number of person and character queries. Our INS approach resembles a prior work focused on interactive retrieval with concept filtering [13]. In our INS implementation, we first compute the mean of each concept score in the S&V database above a secondary threshold (0.10). Other works have sought to dynamically vary this threshold based on distributions, but for the purposes of concept mining and not concept recommendation, this fixed value choice is adequate. Next, we select active concepts for each query image whose individual scores were above each mean score. The active concepts again ranked by their score on each query image and a subset (here only five) of concepts are added to the query as dominant concepts. During query evaluation, we consider only these dominant concepts and compute the Euclidean distance between the query image's and each reference keyframe's concept vector. Keyframes are ordered by their distance and saved for subsequent fusion or verification steps.

3.3 Fusion Strategies

Fusion across multiple modalities has always proven to be difficult for automated search. Systems that achieve the best performance often rely on multiple cues for information, either from multiple instances (i.e. keyframes or textual descriptions) available in the query or through interactive feedback (i.e. relevance feedback). In the INS task, neither of these cues are immediately available, so we turned to an approach that uses query class dependent weightings.

3.3.1 Query class dependency

First applied in previous TRECVID search tasks [14], query class dependent weightings modify their fusion strategies based on the determined class of the query. In traditional search tasks, this approach must automatically select a class based on the components of a query, but in the INS task, the query class is included as part of the query definition. In the INS task this year, the type attribute of each query declares one of four useful query classes: person (8), character (5), object (8), or location (1). Previous studies have shown that the most useful modality for person or character queries is often text whereas object and location

queries are favored by a mix of semantic and low-level feature modalities. This year, we determined the modality weights for each query class by hand because no INS queries were available for automatic training. The weights are shown below in table Table 3. Utilizing these weights, we evaluated

Type	text	dominant	ccd
Person	0.7	0.5	0.3
Character	0.5	0.6	0.1
Location	0.1	0.6	0.9
Object	0.1	0.4	0.8

Table 3: Query class weights for INS task.

each query and each modality and then fused those results at a subshot level to obtain an intermediary result list for subsequent processing.

3.3.2 *Face prioritization*

Harnessing another specific cue from the visual content of each query, face prioritization attempts to assign a higher relevance score to those subshots whose keyframe contains at least one reasonably sized face. For all frames in the S&V reference set, we evaluated the OpenCV face detector [15][16] and stored the coordinates of detected faces as set F. Faces with a pixel area smaller than a certain threshold (here 20x20 pixels) were excluded from consideration as detected regions in F. When applied to a specific run, face prioritization applies a weighting α, such that the relevance score subshot i, denoted as s_i, is boosted if its keyframe contained a face. It should be noted that face prioritization cannot improve the recall of a particular set of results because it is only reranking them.

$$s_i = s_i(1-\alpha) + \begin{cases} \alpha & \text{if } i \in F \\ 0 & \text{otherwise} \end{cases}$$

4. EVALUATION

4.1 CCD Evaluation Results

TRECVID 2010 CCD dataset contains about 12K audio+video query videos, and 12K reference videos. The reference videos are Internet Archive videos, whose quality is much different from those used in TRECVID 2009 CCD task. In total, we extracted 74 millions SIFT features and 110 millions audio features for the reference video set, and 31 millions SIFT features and 63 millions of audio features for the query video set.

This year we submitted 4 runs, labeled as nofa.1 and nofa.3 (for the "no false alarm" profile), and balanced.2 and balanced.4 (for the balanced or equal-error profile). In nofa.1 and balanced.2, we set the visual score weight as 2.0 (audio scores are more dominant), and in nofa.3 and balanced.4, a higher weight of 32.0 is used (visual scores are more dominant).

Overall, our system achieves reasonably good NDCR performance, significantly better than the medium results in all categories. Evaluation results show that runs balanced.1 and balanced.2 perform slightly better than runs balanced.3 and balanced.4. In the rest of this section, we mainly discuss the evaluation results of two runs: nofa.1 and balanced.2. The performance of these two runs are respectively shown in Figure 7 and Figure 8.

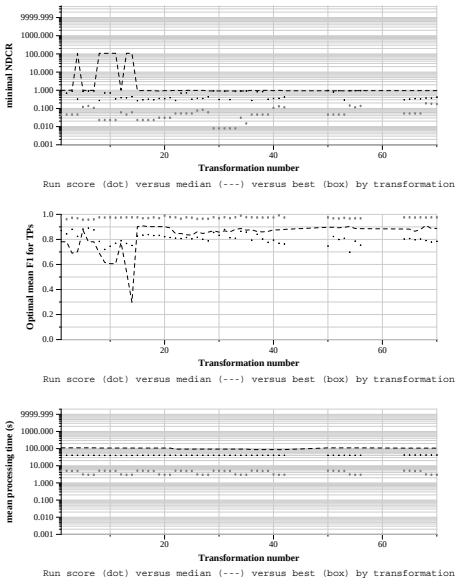

Figure 7: Performance of ATTLabs.NoFA.1

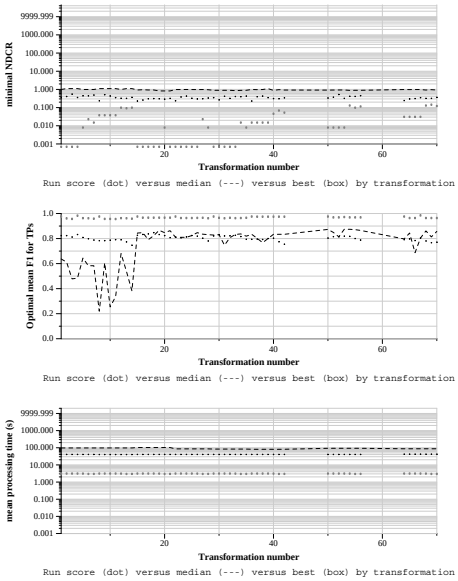

Figure 8: Performance of ATTLabs.Balanced.2

Compared to the results that our system achieved in 2009, the overall relative performance is a bit worse. Possible reasons are 1) the system is over trained for the old evaluation dataset; 2) the audio approach was largely unoptimized; and 3) fusion of audio/visual runs may be affected by spurious frame rates reported by an internal utility. Looking at gains in system speed alone, we achieved significant improvements. In 2009, our speed is much slower than the median speed, yet this year, our speed is faster than the median value. This observation proves that the new hash value indexing and query mechanism works well.

4.2 INS Evaluation Results

The INS task was evaluated on a testing partition of the TRECVID 2009 S&V testing dataset. Using our own segmentation algorithm, from 422 unique videos, a total of

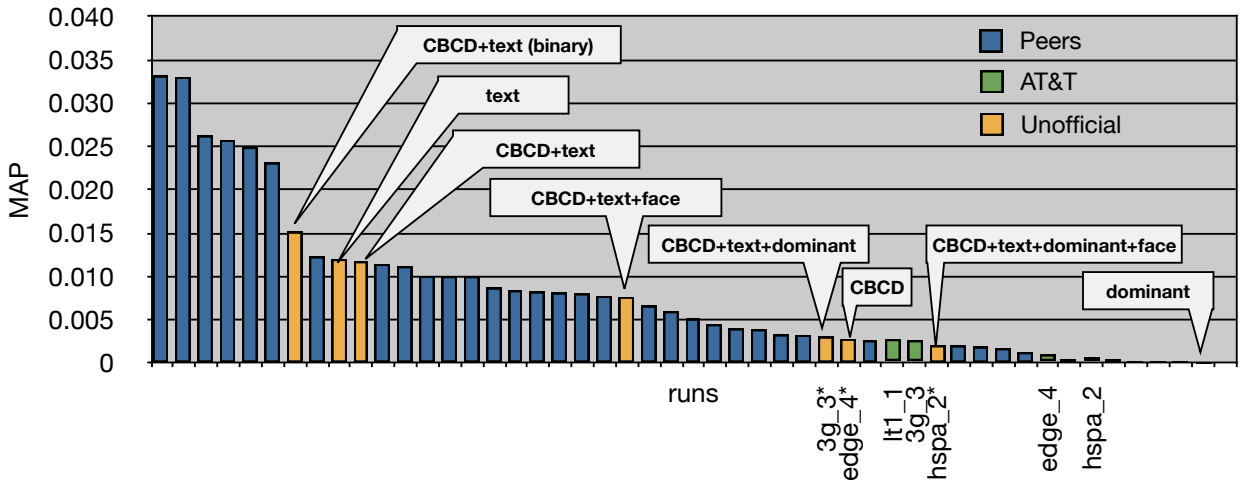

Figure 9: Performance of Automated INS runs. With an asterisk (*), we denote AT&T's official runs after correcting a software error. Unofficial runs for several independent search methods are also illustrated here.

109135 subshots were extracted with an average length of 2.9 seconds. Each of the 22 queries was evaluated with each modality (if available) and subshots were merged to a final list of 1000 shots defined by NIST. Performance, determined by mean average precision (MAP), of all available runs are shown in Figure 9.

4.2.1 Data mapping error

Upon receipt of the ground truth and performance evaluation by NIST, we discovered a software flaw that incorrectly mapped our internal subshots to those used in submission scoring. Subsequently, performance discussions here highlight observations from corrected versions of official runs (denoted with an asterisk) and additional unofficial runs that help to compare and analyze each search method.

4.2.2 Query class effectiveness

In this evaluation of the INS task, utility of the search methods was fairly correlated to certain query types and an approach that tuned differences between query classes was unnecessary. To our surprise, the run with the best performance was based on a text search alone, which utilized machine translation for indexing. For text search alone, only those topics based on people who were famous politicians or celebrities (*George H. Bush, J. P. Balkenende, Prince Bernhard, C. Powell, Midas Dekkers,* etc.) had non-zero performance numbers. Similarly, for the CCD-based runs, only those image queries with distinct logo and unique coloring patterns returned significant performance. The dominant concept search method was the most ineffective and its failure cases are analyzed more in the next section.

Revisiting the utility of query classes, Figure 10 demonstrates that there was little overlap between different search methods. In cases like this, fusion of different search methods only acts to hurt performance, as illustrated by the fusion cases in Figure 9. As an extreme example of the correlation of the search methods and query classes, a run fused in a binary fashion (all text for people topics, all CCD for others) produced the highest possible performance using our

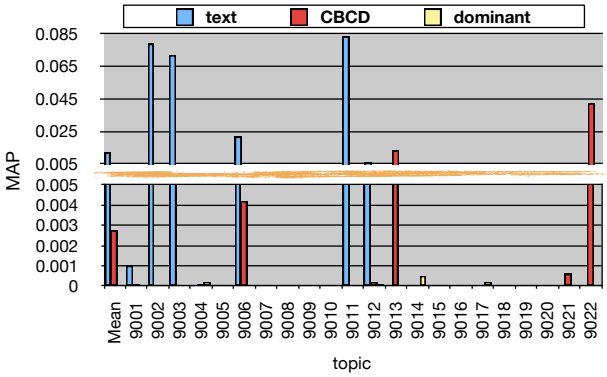

Figure 10: Analysis of Search Method by Topic

search methods.

4.2.3 Face prioritization

Although we did not review the results in great detail, we confirmed that face prioritization was always detrimental for reranking subshots because the detector was just too poor. Often, in person and character-based queries, the image regions containing instances of these people were too small or the persons inside of the search set (S&V 2009 test) were non-frontal face examples, which would immediately downrank any potential hit from another modality.

4.2.4 Analysis of dominant concepts

One problem with the application of semantic classifiers to new domains is the potential for domain shift (i.e. differences in low-level features) to dramatically degrade the accuracy of classifiers. Below, a select set of returned dominant concepts are returned and those particularly relevant are emphasized in Table 4.

While the list of dominant concepts found for each query image has both good and bad recommendations for search, the returned results from a dominant search were often too generic (*person, adult*) or too noisy (*lawn, urban, grand-*

Topic	Id	dominant concepts
George Bush	9001	Furniture, Press Conference, Speaker At Podium, Male News Subject, Dresses Of Women, Politics, Government Leader
J. P. Balkenende	9003	Talking, Sitting, Head And Shoulder, Interview On Location, Observation Tower
office workers	9010	Studio With Anchorperson, Powerplants, Person, Meeting, Adult
Willem Wever van	9022	Computer TV screen, Urban, Lawn, Athlete, Grandstands Bleachers

Table 4: Examples of visually dominant semantic concepts that have high (9001, 9003) and low (9010, 9022) relevance to the query topics.

stands) to accurately select the specific query targets. In future work, we will continue to investigate alternative methods that better leverage the semantic information from the query set as a whole, instead of using each image as an independent query.

5. CONCLUSIONS

In this paper, we reported the AT&T system for TRECVID 2010 evaluation. AT&T participated in two tasks: content-based copy detection and instance-based search. We evaluated numerous alternative representations and enhanced the 2009 SBD system by optimizing indexing and query algorithms. The proposed instance-based search system exploits retrieval techniques from multiple modalities in a fully automated fashion. The evaluation results demonstrate the challenges of the new IACC.1 dataset and difficulty in combining methods for fully automated retrieval systems with minimal query definitions.

6. REFERENCES

[1] Z. Liu, T. Liu, B. Shahraray. "AT&T Research at TRECVID 2009 Content-based Copy Detection." *TRECVID 2009 Workshop*, Gaithersburg, MD, Nov. 16-17, 2009.

[2] A. Smeaton, P. Over, and W. Kraaij. "Evaluation campaigns and TRECVid." In *ACM Multimedia Information Retrieval*, Santa Barbara, California, USA, October 26 - 27, 2006.

[3] L. Fei-Fei, R. Fergus and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories." *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision.* 2004

[4] Everingham, M. Van Gool, L. Williams, C. K. I. , Winn, J. and Zisserman, A. "The PASCAL Visual Object Classes (VOC) Challenge." *International Journal of Computer Vision*, 2010.

[5] M. He ÌĄritier, V. Gupta, L. Gagnon, G. Boulianne, S. Foucher, P. Cardinal. "CRIMâĂŹS CONTENT-BASED COPY DETECTION SYSTEM FOR TRECVID", *TRECVID 2009 Workshop*, Gaithersburg, MD, Nov. 16-17, 2009.

[6] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, and P. Haffner. "A fast, comprehensive shot boundary determination system". In *Proc. of IEEE International Conference on Multimedia and Expo*, 2007.

[7] H. Liu, "MontyLingua: An end-to-end natural language processor with common sense," 2004. Available at: http://web.media.mit.edu/ hugo/montylingua.

[8] "LSCOM Lexicon Definitions and Annotations Version 1.0." In *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia*, Columbia University ADVENT Technical Report #217-2006-3, March 2006.

[9] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. Smeulders. "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia." In *Proceedings of ACM Multimedia*, pp. 421-430, Santa Barbara, USA, October 2006.

[10] B. S. Manjunath, W. -Y. Ma, "Texture Features for Browsing and Retrieval of Image Data", *IEEE Trans on PAMI*, Vol. 18, No. 8, pp. 837-842, 1996.

[11] A. Vailaya, A. Jain, and H.J. Zhang, "On Image Classification: City Images vs. Landscapes", *Pattern Recognition Journal*, Pattern Recognition, Vol. 31, No. 12, pp. 1921-1935, 1998.

[12] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, "SURF: Speeded Up Robust Features", *Computer Vision and Image Understanding* (CVIU), Vol. 110, No. 3, pp. 346–359, 2008.

[13] E. Zavesky, Z. Liu, D. Gibbon, B. Shahraray. "Searching Visual Semantic Spaces with Concept Filters", *IEEE ICSC pp 329-336*, September 17-19 2007.

[14] L. Kennedy, P. Natsev, S.-F. Chang. "Automatic Discovery of Query Class Dependent Models for Multimodal Search." In *ACM Multimedia*, Singapore, November 2005.

[15] Open Source Computer Vision Library, http://www.intel.com/technology/computing/opencv/

[16] P. Viola, and M.J. Jones, "Robust real-time face detection". *Int. Journal of Computer Vision*, 2004.

[17] D. Gibbon, Z. Liu, and B. Shahraray, "The MIRACLE video search engine," IEEE CCNC, Jan. 2006.