AT&T Research at TRECVID 2011

Zhu Liu, Eric Zavesky, Ning Zhou*, Behzad Shahraray

AT&T Labs Research, 200 Laurel Avenue, Middletown, NJ 07748
*University of North Carolina, 9201 University City Blvd, Charlotte, NC 28223
{zliu,ezavesky,behzad}@research.att.com, *nzhou@uncc.edu

ABSTRACT

AT&T participated in two tasks at TRECVID 2011: content-based copy detection (CCD) and instance-based search (INS). The CCD system developed for TRECVID 2010 was enhanced for speed and augmented with an additional picture-in-picture detector and alternative audio features [1]. As a pilot task, participation in INS evaluated object-level content-based copy detection and created a basis for integer-score result reranking. This paper reports the enhancements of the CCD system and briefly describe its application to INS for object-level copy detection.

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 2011 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) known-item search (KIS), multimedia event detection (MED), and instance-based search (INS) as a pilot task. In this paper, systems are described for the CCD and INS tasks and brief initial reactions from the formal TRECVID evaluations are discussed.

Instance-based search is still a pilot task in TRECVID 2011 but this year, it focuses on clipped video segments from the BBC Rushes archive. The video in this archive can be described as raw material from which several dramatic series and travel programs are edited. The major challenge that distinguishes the INS task from traditional search is query formulation for a visual object that is explicitly marked in several images that should very 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 and continued in 2011 to measure retrieval capabilities for the BBC videos.

This paper is organized as follows. Section 2 gives a detailed description of the content-based copy detection system. Section 3 addresses work for fully automated instance based search. Evaluation results from TRECVID 2011 are presented and discussed in Section 4, and conclusions are summarized in Section 5.

2. CONTENT-BASED COPY DETECTION

name	description
att.m.NOFA.1	Combine 2 audio-based detection results and 7 video-based detection results; Fusion weights for audio and video are 1.45 and 0.55; Threshold is 0.23; Only select the top match
att.m.BALANCED.2	Combine 2 audio-based detection results and 7 video-based detection results; Fusion weights for audio and video are 1.4 and 0.6; Threshold is 0.22
att.m.NOFA.3	Combine 4 audio-based detection results and 12 video-based detection results; Fusion weights for audio and video are 1.45 and 0.55; Threshold is 0.22; Only select the top match
att.m.BALANCED.4	Combine 4 audio-based detection results and 12 video-based detection results; Fusion weights for audio and video are 1.4 and 0.6; Threshold is 0.23

Table 1: CCD run names and descriptions.

2.1 Task Overview

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 2011 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

Each original query is expanded to 56 versions of audio+video queries using different combinations of audio and video transformations. In total, 4 runs were submitted for CCD evaluation, 2 in no false alarm profile (att.m.NOFA..1 and att.m.NOFA.3), and 2 in balanced profile (att.m.Balanced.2 and att.m.Balanced.4). Brief descriptions of these runs are listed in Table 1.

2.2 Overview of the CCD System

The main improvements of this year's CCD system are: a new picture-in-picture detector and a new copy match score normalization scheme. Figure 1 illustrates a high level overview of the CCD system. The video-based and audio-based approaches work independently and each module produces a CCD result. The fusion step is a linear weighting with normalization mechanism to combine the audio and video results together. By adjusting the weights, either the audio or the video modality is made more influential on the overall CCD run. Finally, runs tuned for different profiles are created based on these fused results.

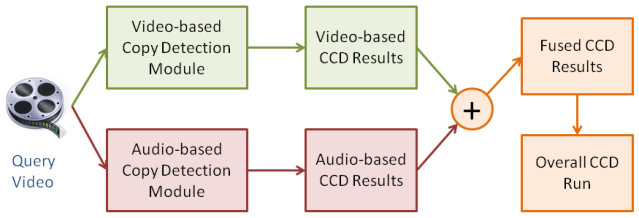


Figure 1: Diagram of the audio/video CCD system

2.3 Video-based CCD sub system

Figure 2 shows the overview diagram of the video-based CCD system. It mainly consists of two parts as indicated by different colors in the figure. The top portion illustrates the processing components for the query videos, and the bottom portion shows the processing stages for the reference videos.

For details of the video processing, please consult a prior notebook paper from TRECVID 2010 [1]. This year, focus was given to two components: 1) improving the transformation detection and normalization module, specifically, the picture in picture detection; and 2) video matching score normalization.

2.3.1 Picture in picture detection

The PiP detection method is essentially based on the observation that the PiP region boundary consistently appears across the whole video. By considering other properties

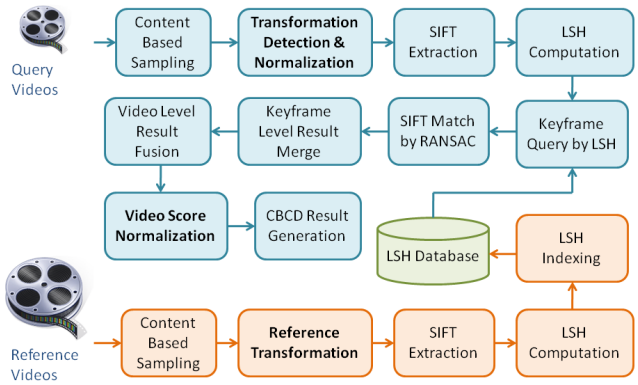


Figure 2: The video-based CCD algorithm

of the PiP region, for example, the geometric constraint, we proposed a detection method as illustrated in Figure 3. Given a video of N frames, at frame f, we detect the PiP region based on all frames from the beginning to frame f. The final result is determined by aggregating the detected regions for the entire video.

Following steps are used to detect the PiP region in current frame.

- Step 1: The image gradient of current frame f (gray image of that frame) is calculated by using the Laplacian filter.
- Step 2: The gradients of frames 0 to f are accumulated and then averaged to form an average gradient image G (a matrix) which is threshold to get a binary mask B using Otsu's algorithm.
- Step 3: The Hough transform is applied on the binary image B to achieve an edge image E of vertical and horizontal edges. The Hough transform is adopted in this context aiming to remove noisy edges. An edge smoothing is also used to get rid of some particular short and long edges.
- Step 4: A similar process as Step 2 is employed to produce a more robust edge image B'.

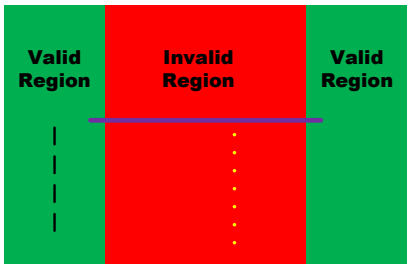


Figure 4: Illustration of the crossing constraint. Given a horizontal edge, the green region is valid for a vertical edge (e.g., the blue dash edge) to appear. The valid region only considers crossing constraints for the current horizontal edge.

- Step 5: Based on the edge image B', multiple candidate rectangles (a PiP region is bounded by a rectan-

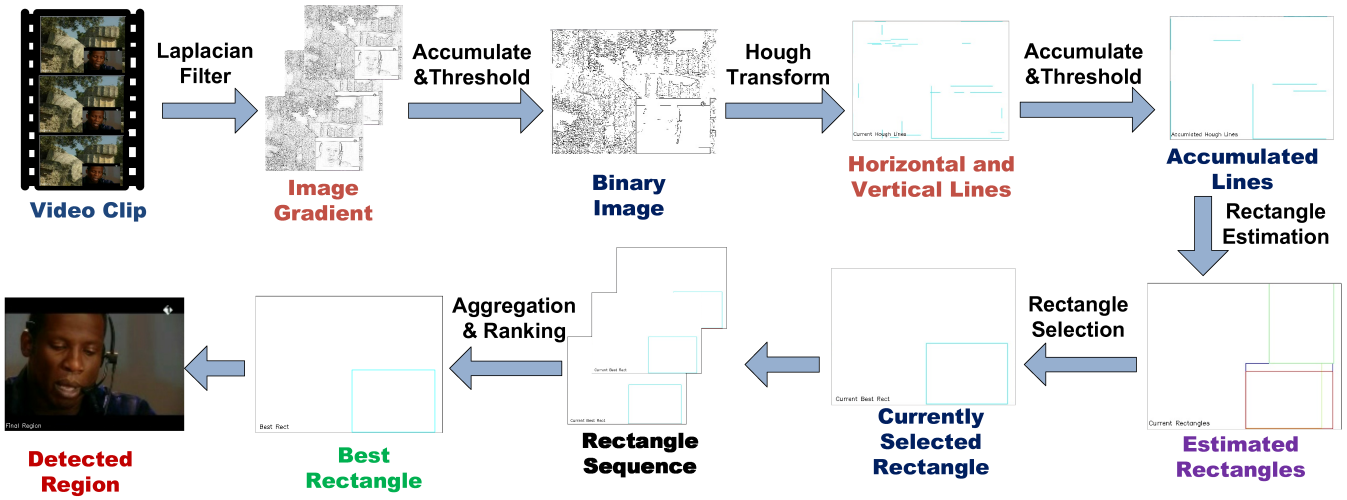


Figure 3: System overview of the picture in picture detection method.

gle) are estimated and assigned with scores by considering the geometric constraints, including the aspect ratio and the size of a rectangle, etc. To remove invalid rectangles, the following criteria is used.

- *Distance constraint*: The parallel edges of a rectangle should not be too close or too far way.
- *Location constraint*: The vertical edges should be located between the two horizontal edges, and the horizontal edges should be located between the two vertical edges.
- *Crossing constraint*: The extended lines of two orthogonal edges of a rectangle cannot cross at the middle of each edge. Figure 4 illustrates this constraint. Given a horizontal edge (the violet solid line in Figure 4), the vertical edges should be placed at the valid region (the green region in Figure 4). For instance, while the blue dash line is a valid vertical edge given the horizontal edge, the yellow dot line is not a valid one.

After removing the invalid rectangles, a valid candidate rectangle set $\mathcal{R} = \{R_i\}_{i=1}^{M}$ is obtained.

- Step 6: For each rectangle R_i, the system computes a score based on its aspect ratio and size and then chooses the rectangle with the highest value as the detected PiP region for the current frame. The score of rectangle R_i is calculated as

$$S(R_i) = S_A(R_i) \times S_S(R_i) \times E(R_i), \tag{1}$$

where $S_A(R_i)$ and $S_S(R_i)$ are the scores of R_i based on its aspect ratio and size respectively, and $E(R_i)$ is the number of detected edges for R_i. $S_A(R_i)$ is essentially based on the agreement between the aspect ratio of R_i and that of the background frame F,

$$S_A(R_i) = \exp\{-|AR(R_i) - AR(F)|\}, \tag{2}$$

where $AR(\cdot)$ is the aspect ratio of a rectangle (width / height). $S_S(R_i)$ is computed as,

$$S_S(R_i) = \begin{cases} 1 & \theta_1 \leq \frac{size(R_i)}{size(F)} \leq \theta_2, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where $size(\cdot)$ computes the area of a rectangle. θ_1 and θ_2 are two predefined thresholds.

At each frame, these six steps estimate one PiP region and after processing the whole video, a set of detected regions P is produced. For a candidate region C_i, a score is assigned based to two strategies. One strategy is by voting, and the other is by summation of scores. In both cases, the top ranked detected region is chosen as the final PiP result.

- *Voting.* The score of each candidate region C_i is the number of frames within which C_i is detected as a PiP region.
- *Summation.* This strategy uses the summation of scores of C_i in each frame within which it is detected as a PiP region as the final score of region C_i.

2.3.2 *Score normalization*

The score normalization is inspired by the method reported in [6]. Assuming that a query video has N matches, and their original video matching scores are $s'_v(i), i = 0, ..., N-1$, the normalized scores $s_v(i)$ are computed by the following formula,

$$s_v(i) = \alpha(i) \times sigmoid(s'_v(i)), \ where \tag{4}$$

$$\alpha(i) = \begin{cases} \frac{M-i}{M} \times \frac{s'_v(i)}{\sum_{j=0}^{M-1} s'_v(j)} & \text{for } \theta_1 \leq i < M \\ \alpha(M-1) & \text{for } i \geq M \end{cases} \tag{5}$$

M is a preset number, set to 8 in this system. When N is smaller than M, the last match score is repeated and the number of match videos is extended to M. As shown in the formula, the weight $\alpha(i)$ is determined by both the rank, i, and the match score $s'_v(i)$. Different from the method in [6], the original match score is processed by a sigmoid function. The range of the final normalized scores is $[0, 1)$.

2.3.3 *Query normalization and reference transformation*

For each query keyframe, PiP and letterbox regions are detected. The detected PiP region is resized to the half

size of the original keyframe. The detected letterbox region is removed, and the rest of the content is shifted and scaled to the original size. In addition, the query keyframe is equalized and blurred to overcome the effect of change of Gamma and white noise transformations. To deal with the flip transformation, a flipped version is created for each of these normalized query keyframes. In summary, there are 10 types of query keyframes: original, letterbox removed, PiP scaled, equalized, blurred, and flipped versions of these five types.

Complementary to normalizing the query keyframes, the reference keyframes can be pre-processed to eliminate the transformation effect as much as possible. In this work, the system simply applies two transformations on reference keyframes. They are half resolution rescaling and strong re-encoding (using ImageMagick with the quality parameter set to 10). In total, there are three different types of reference keyframes. As shown in Table 2, the system in this work considers 12 combinations of normalized query keyframes and transformed reference keyframes.

Pair	Query keyframes	Reference keyframes
1	Original	Original
2	Flipped	
3	Letterbox removed	
4	Letterbox removed and flipped	
5	Equalized	
6	Equalized and flipped	
7	Blurred	
8	Blurred and flipped	
9	Original	Encoded
10	Flipped	
11	Picture in Picture (PiP)	Half
12	PiP and flipped	

Table 2: Normalized query keyframes vs. transformed reference keyframes.

2.4 Audio-based CCD sub system

The audio based CCD system is similar to the approach reported in TRECVID 2010. Figure 5 shows the diagram of the audio-based approach. For the reference audios, the system first creates a set of transformed versions to cope with the audio transformations in the queries. The transformations considered are 1) compression, 2) bandpass filtering, and 3) companding. The reference audio is compressed using ffmpeg in MP3 format with a bit rate of 16 kpbs. The bandpassed audio is created by sox with a frequency bandwidth of 500 Hz to 3000 Hz. The companding transformation is also generated by the sox tool. Unlike the video-based approach, transformation detections are not performed on the query audio. In total, there are four pairs of query-reference audio transforms: 1) original query vs. original reference; 2) original query vs. compressed reference; 3) original query vs. bandpass reference; and 4) original query vs. companding reference.

For the original and each transformed reference audio, fingerprints are computed and indexed using energy difference between the sub-bands for each frame. In this work, the original audio signal is re-sampled in 16 KHz. Each frame is 32 milliseconds long, and adjacent frames overlap by 22 milliseconds (which leads to 100 frames per second). Our

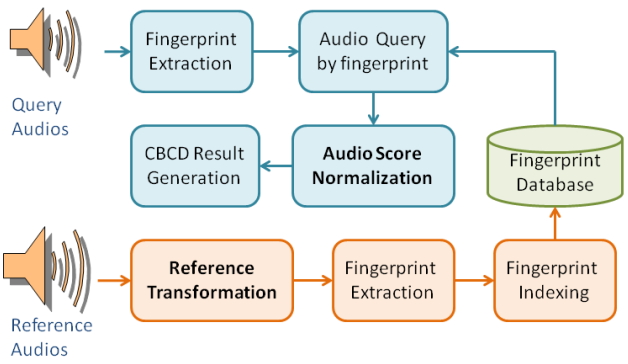


Figure 5: The audio-based CCD algorithm.

system considers 17 subbands and produces a 16 bit fingerprint, where the 17 subbands are in Bark scale between 300 Hz and 7700 Hz.

For each query audio, the system computes the fingerprint and then searches the reference fingerprint database to find matching reference audios. The audio matching scores are normalized in the same way described in the video-based approach.

2.5 Audio-Video Fusion

Audio and video based CCD results are fused at the final stage. For each query, both audio-based and video-based CCD modules report a list of matches. Each match is specified by a query segment, a reference segment, and the associated matching score. When the two lists are merged, if both the query and reference segments of an audio-based match overlap with those of a video-based match respectively, and the overlapped regions are more than half of the original segments, then the two matches are merged and the merged segment is inserted in the fused match list. Otherwise, the original audio (or video) based match is copied to the fused match list with the score weighted by a factor w_a (or w_v). When overlapped matches are merged, the new query/reference segment is the union of the two original query/reference segments, and the score is a weighted sum of the original scores: w_a for the audio-based match score and w_v for the video-based match score.

After merging, the fused match list is sorted based on new scores, and then normalized in the same way described in video-based sub system. While generating runs for the no false alarm profile, only the best matches whose scores are higher than a certain threshold are reported. For the balanced profile, all matches that are higher than certain threshold are reported.

3. INSTANCE-BASED SEARCH

3.1 Task Overview

The instance-based search task in TRECVID 2011 permitted an object-level evaluation of the content-based copy detection system presented above. Contrary to work last year [1], the system formulated for INS evaluation this year relied almost exclusively on the content-based copy detection algorithm. This modification was made because scene level semantics (i.e. semantic concepts or textual annota-

tions) and person-based re-ranking (i.e. using face detection) yielded worse performance in TRECVID2010.

3.2 Instance Search Methods

Figure 6 illustrates the INS pipeline. Leveraging the main CCD algorithm, the major contributions of the INS framework are the use of region-based re-ranking and score normalization for correspondence to the top-N search formulation of the INS task. As stated above, this year's approach

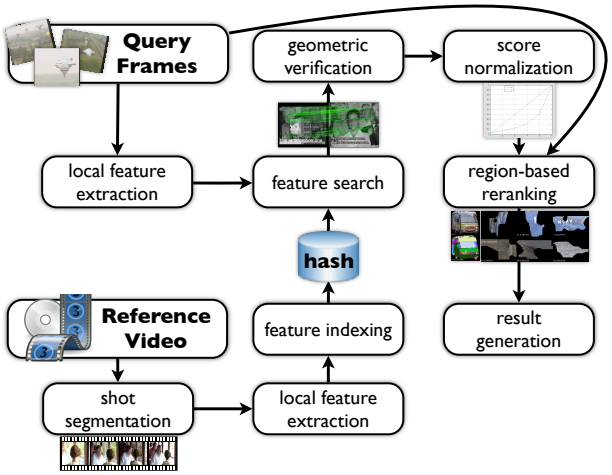


Figure 6: Overview of INS system from reference indexing to query evaluation (feature search with verification and result re-ranking).

for the INS task was heavily focused on matching visual content at an object-level. Instead of combining scene level semantics, the approach in Figure 6 relies on content-based copy detection techniques (Section 2.3) to first pool a set of candidate results. Then, a score normalization is performed to combine the results from multiple (independently evaluated) query frames. Finally, visual information from a region spatially limited to the query (Section 3.2.2) is utilized to re-rank raw CCD results.

3.2.1 *Object queries with context*

An important consideration when approaching an object-level search problem is whether or not to include the context of the object in the query. The video component of the content-based method (Section 2.3) utilizes local features (i.e. SURF [8]) to index the objects and scenes in a video keyframe. While there is no definitive choice between sparse sampling (i.e. interest point detectors) or dense sampling of a scene for local visual features ([9], [10]). Dense sampling has greater resource demands because it returns more samples, takes more time to compute, and may reduce the robustness of geometric verification due to imprecise feature localization as it utilizes grid coordinates rather than visually salient ones. The shortcoming of sparse detection is that smaller objects and those that have a homogenous and non-textured appearance may not be represented by the features. Figure 7 illustrates objects that are unaffected like 7(a) and 7(b) (query 9035, "tortoise") and those objects that are likely missed like 7(c) and 7(d) (query 9036, "all yellow balloon") with sparse sampling. For these reasons, images processed as queries by this year's INS system only utilize

(a) (b) (c) (d)

Figure 7: Images for (a) query 9035 and (c) query 9036 and their respective masks (b) and (d). Some objects (query 9036) may be too small for sparse sampling to index without scene context.

the mask information provided with the query (7(b) and 7(d)) during secondary re-ranking stages (section 3.2.2).

3.2.2 *Region-based re-ranking*

Region-based re-ranking is a method that emphasizes local similarity of image patches as illustrated below. First, an image is analyzed and segmented (Figure 8(a)) into non-exclusive regions that may overlap. The algorithm used in this work for segmentation was MSER (maximally stable extremal regions [13]). Next, patch candidates that reside within the masked query area are used to match reference images. Here, the matching stage uses a weighted maximum similarity between low-level color features and Hu's set of invariant moments (Figure 8(b)). Given that the initial visual CCD results are integer-based (i.e. the number of matching pairs), preliminary results also indicate a slight advantage for re-ranking of results only with the same integer score. For example if a set of CCD reference-score pairs is $S_{ccd} = \{(x_0, 6), (x_1, 5), (x_2, 5), (x_3, 4), (x_4, 4), (x_5, 4)\}$ then region-based re-ranking would occur on three discrete sets: $S^1_{ccd} = \{x_0\}, S^2_{ccd} = \{x_1, x_2\}, S^3_{ccd} = \{x_3, x_4, x_5\}$.

(a) (b)

Figure 8: Demonstration of (a) MSER region segmentation and (b) region similarity scoring with color grid moments and Hu's invariant moments.

As illustrated in Figure 8, the proposed method is sensitive to dramatic shape or color shifts, which can produce artificially high similarity scores. Continuing investigation is focused on deriving stable appearance features that capture both color and shape such that automatically extracted MSER regions are accurately matched between reference and query images.

3.2.3 *Score normalization*

Although intermediary scores of the INS results have similar distributions to those in the the CCD results, an alternate normalization technique is defined here. The normalization method defined here is applied across an entire list of results (to accommodate the top 1000 result requirement) contrary to the method defined in Section 2.3.2, where only the top M results are modified (even though M in Equation 5 can be arbitrarily large).

Generally, scores from the CCD method above can be characterized as homogenous or heterogenous by analyzing the top M results, as shown in Figure 9. The normalization

Figure 9: Query image results exhibiting a (top) hetereneous and (bottom) homogenous score distribution.

process defined in this section is applied so that score results from each query image should be more amenable to averaging with other query images for the same topic. First, if a query frame has N matches and its original matching scores are $s(i), i = 0, ..., N-1$, range normalized from $[0, 1]$,

$$s_r(i) = \frac{s(i) - s_{min}}{s^{max} - s^{min}}. \quad (6)$$

are as $s_r(i)$. With these range normalized scores, determine the mean $\bar{s_r}$ and median $\tilde{s_r}$ values from the top M results (i.e. those with the highest matching score). In this work M is set to 10. As determined by these two values, α is determined to indicate distribution shape,

$$\alpha = \begin{cases} 1 - \tilde{s_r} & \text{if } \bar{s_r} < \tilde{s_r} \\ \tilde{s_r} - 2 & \text{otherwise} \end{cases}. \quad (7)$$

such that a concave exponential is applied homogenous score sets to emphasize minute changes and a convex exponential is applied to heterogeneous sets to dampen differences and better utilize the score interval. Finally, a transformed score set $s_t(i)$ is derived,

$$s_t(i) = \frac{\alpha s_r(i)}{\alpha - s_r(i) + 1}. \quad (8)$$

from the range normalized scores $s_r(i)$. Subsequent averaging between query frames operates on these transformed scores.

4. EVALUATION

4.1 PiP Evaluation Results

Using the query videos with PiP transformation from TREVCID 2009 and 2010, this section assesses the performance of the proposed PiP detector. Both TRECVID 2009 and 2010 datasets contain 201 query videos with PiP transformation, while the PiP regions in TRECVID 2010 dataset are created more dynamically in terms of their sizes and locations and the detection is more challenging.

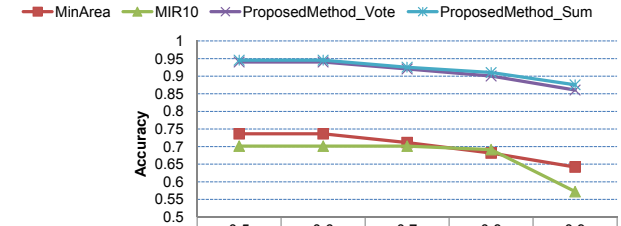
The agreement between the area of the detected PiP region and that of the ground truth is used as the evaluation metric. In particular, let O be the overlapping area of the detected PiP region R and the ground truth PiP region S. The precision and recall rates are defined as,

$$\text{precision}(R) = \frac{\text{size}(O)}{\text{size}(R)}; \ \text{recall}(R) = \frac{\text{size}(O)}{\text{size}(S)},$$

where $size(\cdot)$ is a the area of a rectangle. The PiP region is successfully detected if both the recall and precision rates are higher than a threshold T. The detection accuracy is calculated as,

$$\text{accuracy} = \frac{\text{Number of successful PiP detection}}{\text{Number of test videos}}.$$

To assess the false alarm rate of the proposed method, a dataset with the flip transformation from the TRECVID 2009 dataset is also prepared. To be precise, video clips in the "TRECVID 2009 flip dataset" are absolutely free of the PiP transformation. The two baseline methods are MinArea

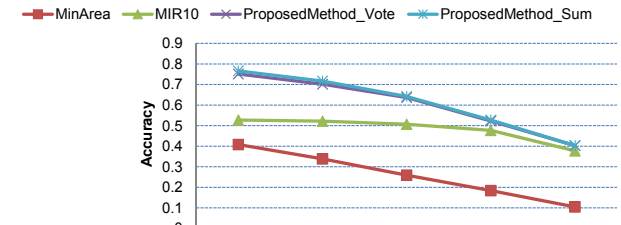


	0.5	0.6	0.7	0.8	0.9
MinArea	0.736318	0.736318	0.711443	0.681592	0.641791
MIR10	0.701493	0.701493	0.701493	0.691542	0.572139
ProposedMethod_Vote	0.940299	0.940299	0.920398	0.900498	0.860697
ProposedMethod_Sum	0.945274	0.945274	0.925373	0.910448	0.875622

Threshold

Figure 10: PiP detection performance on the TRECVID 2009 PiP dataset.

method and the method used in [1] denoted as MIR10 in the sequel. The MinArea method follows the first four steps of the proposed method but detect the PiP region by finding the minimum area rotated rectangle which covers all the endpoints of the edges in the edge image B'. The MIR10 approach accumulated the edge detection information across the whole video and computed the PiP region by projecting the accumulated edge evidence onto horizontal and vertical edge profiles. For more details, please refer to [1]. Figure 10 shows the comparison on the TRECVID 2009 dataset. Accuracy is demonstrated with the threshold T ranging from 0.5 to 0.9. It is clearly seen that the proposed method exhibits a significant performance boost. Compared with MIR10 method, the accuracy gain of the proposed method using summation aggregation strategy ($ProposedMethod_Sum$ in Figure 7) is 31.65% when the threshold is 0.8. Figure 11



	0.5	0.6	0.7	0.8	0.9
MinArea	0.40796	0.338308	0.258706	0.18408	0.104478
MIR10	0.526866	0.521891	0.506965	0.477114	0.377612
ProposedMethod_Vote	0.751244	0.701493	0.636816	0.522388	0.402985
ProposedMethod_Sum	0.766169	0.716418	0.641791	0.527363	0.402985

Threshold

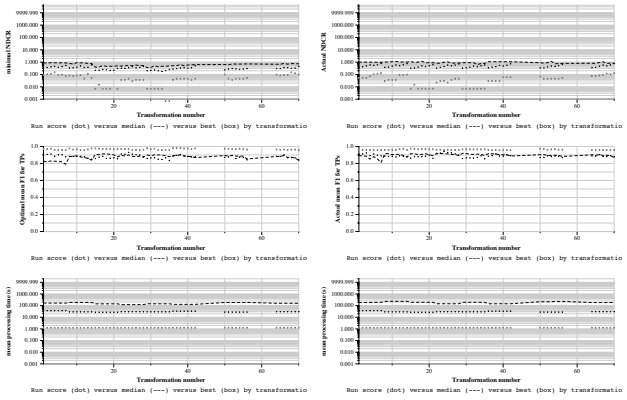
Figure 11: PiP detection performance on the TRECVID 2010 PiP dataset.

presents an analysis on the TRECVID 2010 PiP dataset. Under all the threshold settings, similar accuracy boosting can be observed. The TREVID 2009 flip dataset contains 201 query videos, among which 31 videos have been detected as having a PiP region by the proposed method with summation aggregation strategy. The false alarm rate is estimated as 15.43%.

4.2 CCD Evaluation Results

TRECVID 2011 CCD dataset contains about 12K audio+video query videos, and 12K reference videos. In total, 82 million SIFT features and 110 million audio features were extracted for the reference set, and 33 million SIFT features and 63 million audio features for the query set.

This year, 4 runs were submitted: att.m.NOFA.1 and att.m.NOFA.3 (for the NoFA profile), and att.m.BALANCED.2 and att.m.BALANCED.4 (for the balanced profile). In runs att.m.NOFA.1 and att.m.BALANCED.2, a total of 7 combinations of video transformations (pairs 1, 2, 3, 8, 9, 10, and 11 listed in Table 2) and 2 combinations of audio transformations (original query vs. original reference, and original query vs. compressed reference) were used. The choice of these combinations is determined based on the CCD performance on TRECVID 2010 dataset. For the other two runs, all 12 combinations of video transformations and all 4 combinations of audio transformations were used. For runs in the no false alarm profiles, the audio score weight w_a as 1.45 and video weight w_v to 0.55, and for runs in the balanced profiles, these weights are set to 1.4 and 0.6. Thresholds are set to 0.23 for runs att.m.NOFA.1 and att.m.BALANCED.4, and 0.22 for the other two runs. These parameters are determined based on the TRECVID 2010 dataset.



(a) (b)

Figure 12: Performance of (a) ATTLabs.NoFA.1 and (b)ATTLabs.Balanced.2

Overall, this system achieves reasonably good NDCR performance, significantly better than the medium results in all categories. Evaluation results show that run att.m.BALANCED.2 performs slightly better than run att.m.BALANCED.4, and runs att.m.NOFA.1 and att.m.NOFA.3 perform similarly. The remainder of this discussion focuses on the evaluation results of two runs: att.m.NOFA.1 and att.m.BALANCED.2. The performance of these two runs is shown in Figure 12.

Compared to the results achieved in 2010, the overall performance is much better. For example, the average optimal NDCR for the balanced profile is reduced from 0.47 to 0.32. It is obvious that the new match score normalization scheme is very effective. The optimal NDCR for TV2 (PiP transformation) is reduced from 0.66 to 0.42 for the balanced profile, a more significant improvement than the average NDCR. This means that the new PiP detector further boosts the CCD performance. Speed-wise, improvements were also achieved, mainly due to the optimized hash/fingerprint computation, indexing and retrieval. A binary file format is adopted for the entire system in 2011, which significantly increases the file access speed, compared to the system in 2010.

4.3 INS Evaluation Results

The INS task was evaluated on a the TRECVID 2011 BBC rushes dataset. Using a previously published segmentation algorithm [7], a total of 109135 subshots were extracted from 84128 unique video segments from NIST. Each of the 25 queries and its query frames were evaluated independently and then merged to a final list of 1000 videos segments from those defined by NIST.

4.3.1 Result offset error

Upon receipt of the ground truth and performance evaluation by NIST, an index-offset flaw was discovered that offset all mapped reference videos from the system by 1. Consequently, to reproduce results for the the discussion below, the submitted file should use a reference id offset of +1.

4.3.2 Region-based re-ranking

Unfortunately, not enough time to sufficiently evaluate this algorithm for this paper. Future revisions should have more complete results.

4.3.3 Submission analysis

Figure 13 illustrates the performance in mAP (mean average precision) of the submitted run with the maximum and median per-topic scores from all 36 INS submissions. It should be noted that the submitted run was a baseline run relying only on CCD output and score normalization and it does not include region-based re-ranking. Two facts are immediately visible from this figure comparing the submitted run to other submissions: (1) the system described above is generally competitive and better than the median score and (2) for the most part, the submitted run generally follows

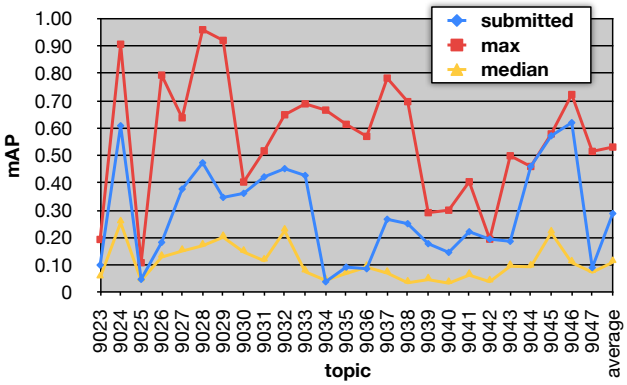


Figure 13: Performance of INS runs in mAP (mean average precision) with max and median over all submissions.

the high and low performance trends of its peers. Given these observations, the remainder of this section focuses on the inconsistencies or errors with respect to these trends.

Looking at outlier topics, 9034, 9035, 9036, 9047 resulted in unusually low performance. Upon closer analysis of these query topics, these topics can be grouped into three failure categories: homogenous object appearances, inconsistent object visuals, and objects being too small with respect to their scene. First, topics 9034 ("tall, cylindrical building") and 9035 ("tortoise", Figure 7(a)) exhibit a fairly homogenous object appearance. Here, a homogenous appearance implies a generally low-contrast texture or surface and one that possesses few distinct visual features (i.e. edges, sharp lines, corners, etc.). This weakness is known for local feature representations and possible solutions may incorporate more contextual information (looking into the scene for cues) or alternative feature representations that capture region or shape size (i.e. joint MSER segmentation and local feature representation). Second the objects in query topic 9036 ("all yellow balloon", Figure 7(c)) and 9047 ("airplane-shaped balloon") demonstrate cases where the object's visual features are overwhelmed by the context of the query frame during indexing. Typically this problem is exacerbated when the object is too small, as in topic topic 9036, but indexing with independent visual features can penalize results that have large green fields or expansive blue skies by incorrectly retrieving unrelated scenes with these large regions. Future implementations will revisit the inclusion of contextual features (Section 3.2.1) due to these potential negative side effects.

Next, a number of topics 9042, 9044, 9045 resulted in unusually high performing topics. Topics 9044 ("American flag") and 9045 ("lantern") both define very small objects within a fairly complex scene context. This observation is particularly interesting because it demonstrates two cases where contextual features (Section 3.2.1) actually improved search performance, whereas the last paragraph discussed two topics that were harmed by these features. A deeper inspection of the reference videos indicates that these topics may have benefited from scene context because the relevant reference videos for topics 9044 and 9045 often contained prolonged shots of this scene. Not surprisingly, the reference videos for topics 9036 and 9047 (see discussion above) contained fewer long-duration or consistent appearance (i.e. steady camera) shots. While additional investigation is required, these findings may help to explain the unusually good and bad performance of topics exhibiting roughly the same query characteristics. Finally, topic 9042 ("male presenter Y") is also grouped with those aided by scene context even though its query object (a person) is relatively large because at the time of writing no better explanation can be reached.

The TRECVID2011 INS evaluation demonstrates that the CCD system provides a good baseline for object-level search. Some questions remain open regarding the qualitative benefits of score normalization, region-based re-ranking, and the use of scene context (as opposed to object visual alone). However, given the relative performance of this baseline system and other peers in the evaluation, a fair conclusion may be that future challenges will come from the nature of the instance queries (incorporating motion, events, dependent objects, etc.) and the reference content (permutations of the copied content, homogenous surveillance-like sources, etc.).

5. CONCLUSIONS

This paper documents the AT&T system and results for the TRECVID 2011 evaluation. AT&T participated in two tasks: content-based copy detection (CCD) and instance-based search (INS). For the CCD task, innovations focused on the picture-in-picture detection module and a copy match score normalization scheme. The evaluation results show the effectiveness of these improvements. The proposed instance-based search system was modified this year to leverage the indexing robustness of the CCD system to perform object-level retrieval.

6. REFERENCES

[1] Z. Liu, E. Zavesky, N. Sawant, B. Shahraray. "AT&T Research at TRECVID 2010." *TRECVID 2010 Workshop*, Gaithersburg, MD, Nov. 15-17, 2010.

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

[5] M. Héritier, V. Gupta, L. Gagnon, G. Boulianne, S. Foucher, P. Cardinal. "CRIM's Content-Based Copy Detection System for TRECVID", *TRECVID 2009 Workshop*, Gaithersburg, MD, Nov. 16-17, 2009.

[6] E. Younessian, X. Anguera, T. Adamek, N. Oliver, and D. Marimon, "Telefonica Research at TRECVID 2010 Content-based Copy Detection," TRECVID 2010, Gaithersburg, MD, 2010.

[7] 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.

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

[9] Y.-G. Jiang, J. Yang, C.-W. Ngo, A. Hauptmann, "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study", *IEEE Transactions on Multimedia*, vol. 12, issue 1, pp. 42-53, 2010.

[10] K. van de Sande, T. Gevers, C. Snoek, "Evaluating Color Descriptors for Object and Scene Recognition", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 32 (9), pages 1582-1596, 2010.

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

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

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. of British Machine Vision Conference*, 2002, pp. 384-393.