NTT Communication Science Laboratories at TRECVID 2011 Content-Based Copy Detection

Ryo Mukai, Takayuki Kurozumi, Takahito Kawanishi
Hidehisa Nagano, Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa 243-0198, Japan
{mukai.ryo, kurozumi.takayuki, kawanishi.takahito,
nagano.hidehisa, kashino.kunio}@lab.ntt.co.jp

Abstract

In this paper, we describe our approaches that were tested in the TRECVID 2011 Content-Based Copy Detection (CBCD) task. We use our fingerprinting technologies called the Coarsely-quantized Area Matching method (CAM) and Divide And Locate method (DAL) for video detection and audio detection tasks, respectively. The CAM consists of a feature degeneration and sparse feature selection process. The DAL is based on spectral partitioning and vector quantization. The audio and video are processed independently, and we merged the audio and video results to generate submission runs.

We submitted 4 runs with varying combinations of audio and video search engine settings. The features and search parameters for intermediate search results are common to all the runs. The intermediate results are processed with a time consistency filter. We varied the filter parameters as follows.

NTT-CSL.m.nofa.0 : strict filter for both audio and video

NTT-CSL.m.balanced.1 : strict filter for video and weak for audio

NTT-CSL.m.balanced.2 : weak filter for video and strict for audio

NTT-CSL.m.balanced.3 : weak filter for both audio and video

1 System Overview

Figure 1 shows the flow of our copy-detection system configured for the TRECVID 2011 CBCD task. The transformation in the TRECVID CBCD task includes many complex patterns, and we introduced a pre-processing stage in the query video feature extraction.

The reference audio and video data are converted into feature data and stored in a reference database. On the other hand, query video data are pre-processed by several transformations such as cropping, flip, and anti-frame drop. The extracted audio and video features are processed in an audio search engine and a video search engine, respectively. In the search stage, the query features are divided into very short segments and processed with very a low score threshold parameter. Accordingly we obtain massive search result candidates as intermediate results. The intermediate results are processed by the time consistency filter described in Sec. 4. Finally, we merge the audio and video results to generate a submission run.

2 Video Copy Detection

We adopted the Coarsely-quantized Area Matching method (CAM) for the video detection task [1, 2]. The basic algorithm is the same as that used in TRECVID 2010, which is detailed in [3]. This section briefly reviews feature extraction with the CAM.

2.1 Video feature extraction

The CAM uses the color values of each pixel in the video stream. Each frame of the video is divided into small regions, and then the areas in which color changes greatly over time are selected. The feature of the CAM is the time series of the coarsely quantized pixel values of the selected areas.

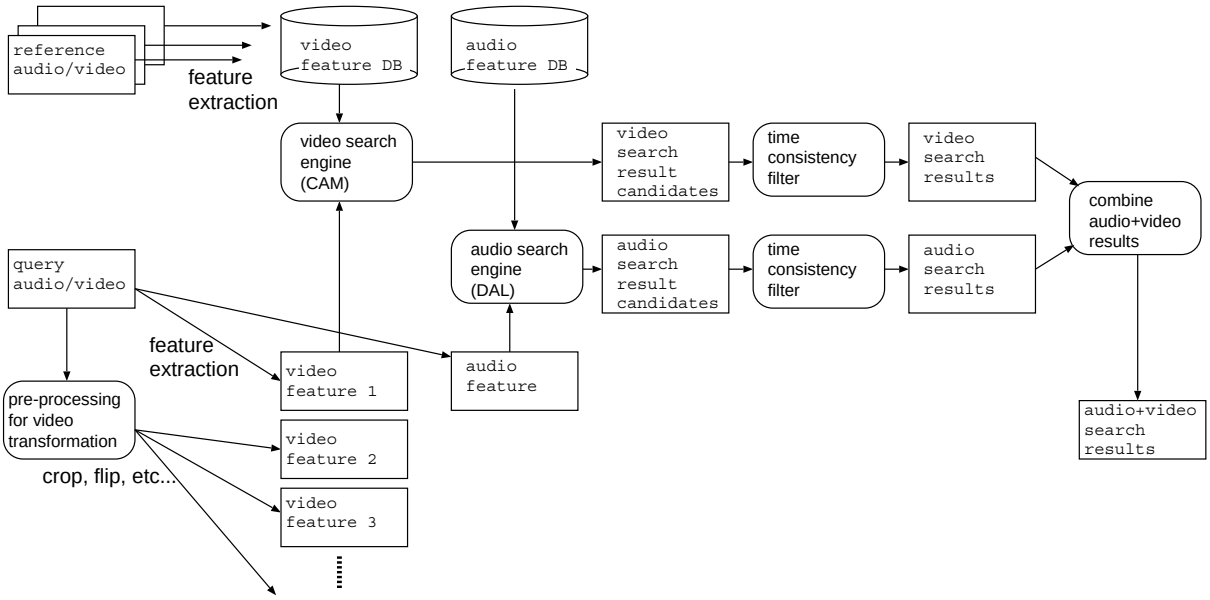

Figure 1: System Overview

Figure 2 shows feature extraction procedure. First, the input video frame is divided into small regions. We adopt the average RGB values over each region as a primitive video feature. On the other hand, we calculate an average and a standard deviation of the pixel values over a time window. We assume that an area that deviates greatly from a temporal local average is a salient part. Based of this assumption, we make a mask for feature selection based on the standard deviations. Next, we quantize the selected feature values. The quantization is carried out on locally normalized feature values. Accordingly, we obtain the feature data of the CAM.

2.2 Pre-processing for multiple feature generation

The conventional CAM is weak against geometrical transformations such as picture in picture type 1 (T2, the original video is small) or flip, since the feature of the CAM is dependent on the positions of the pixel values. Frame dropping, or the insertion of a black/white frame, also harms the CAM feature, because it causes a rapid change of intensity in the temporal direction and CAM mistakes it for a feature.

To cope with video transformations in the TRECVID copy detection task, we introduced a pre-processing stage including several transformations for generating multiple features. The stage was inserted before the query video feature extraction.

3 Audio Copy Detection

We adopted the Divide And Locate method (DAL) for the audio detection task [4]. The basic algorithm is the same as that used in TRECVID 2010, which is detailed in [3]. This section briefly reviews the DAL.

3.1 Audio feature extraction

The basic idea of the DAL is to divide a spectrogram into a number of small regions and undertake matching for each region to locate it in the database. Figure 3 shows the feature extraction of the DAL.

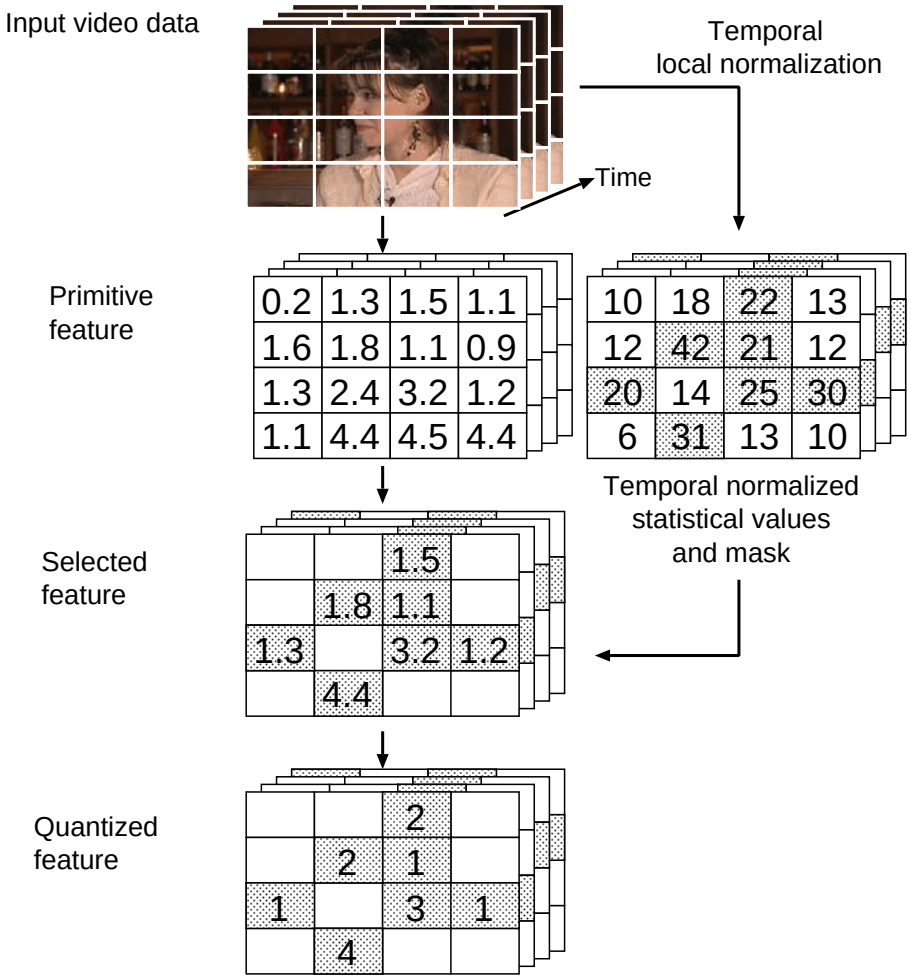

Figure 2: Video feature of CAM

The small spectrum components are quantized by vector quantization (VQ) using codebooks prepared for each frequency band using the LBG algorithm.

3.2 Feature matching

The matching operation is executed by looking up a similarity table among the VQ codes and scanning index lists. This operation is executed much more efficiently than an exhaustive search, therefore the DAL realizes a very fast search over a huge database.

4 Time Consistency Filtering

Last year, in the TRECVID 2010 CBCD task, we found that queries containing short segments to be detected (e.g. less than 10 seconds) tended to be missed and resulted in false negatives. We adopted the following approaches for detecting such short segments.

- very short query window (audio: 0.25 sec. video: 1.0, 2.0, and 3.0 sec.)
- very low threshold

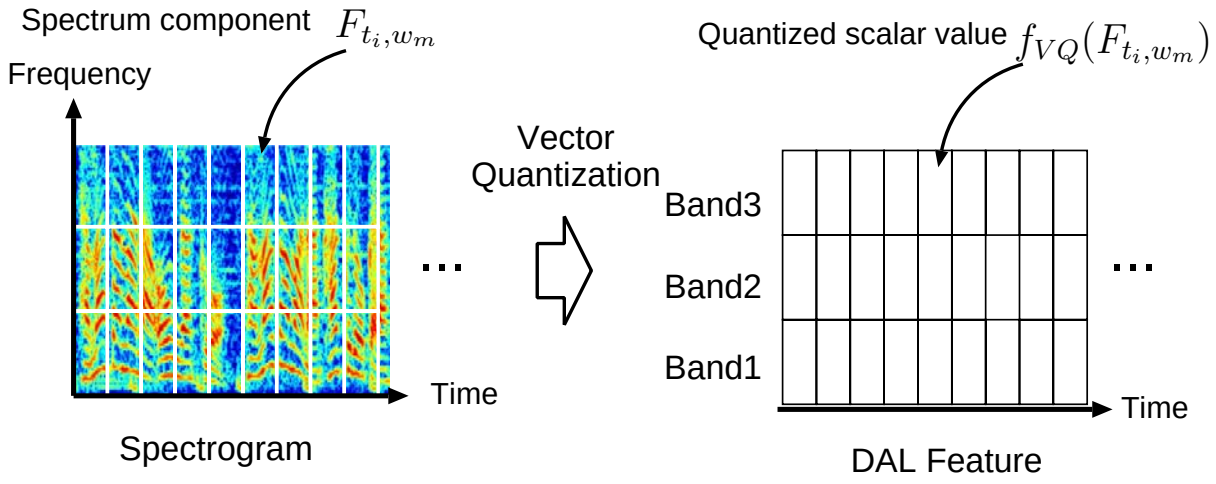

Figure 3: Audio feature of DAL

- time consistency filtering

The TRECVID CBCD 2011 task does not contain a transformation that changes the playback speed, therefore we adopted a very simple algorithm for the filter.

First, we divide the query feature into small segments with a very short window. The query segments are processed by the search engine with a very low score threshold. This gives a large number of (e.g. $N \approx 10000$) candidate results,

$$R_i = (x_i, y_i, ID_i) \quad (i = 1, ..., N), \tag{1}$$

where x_i is the query start time, y_i is the reference start time, and ID_i is a reference ID. The number of time consistent results regarding R_i is calculated by

$$count(R_i) = |\{(x_j, y_j, ID_j) \,:\, x_i - x_j = y_i - y_j \,\text{and}\, ID_i = ID_j \,(1 \le j \le N)\}| \tag{2}$$

Then, we accept the results where $count(R_i) \ge C_{req}$ and obtain a final result as a concatenation of the accepted results. We varied the required number of the count C_{req} as a parameter for the submission runs.

5 TV2011 submissions and results

This section describes the runs we submitted for the TRECVID 2011 CBCD task.

5.1 Audio+Video results

In 2011, audio only and video only results were not tested, but audio + video results were required to be submitted. We merged the audio and video results using the following procedure for each query. We empirically prioritized the audio result when the audio and video results conflicted.

1. If there are overlapping audio and video results with same reference ID, the audio results are accepted.
2. If there is no overlapping result, the audio results are accepted.
3. If there is no audio result, the video results are accepted.

When multiple results overlap on the same query segment, we accept only the top result as regards the length of the detected segment to avoid false alarms.

Table 1: Search algorithms and settings

	algorithm	time consistency filter configuration
audio.BALANCED	DAL	$C_{req} = 10$, screening threshold = 0.9
audio.NOFA	DAL	$C_{req} = 15$, screening threshold = 0.94
video.BALANCED	CAM	$C_{req} = 5$, screening threshold = 0.7
video.NOFA	CAM	$C_{req} = 5$, screening threshold = 0.8

5.2 Submitted results

We submitted four runs with varying time consistency filter settings. These are summarized in Table 1. The labels and combinations of the settings of the submitted runs are as follows

NTT-CSL.m.nofa.0 : video.NOFA + audio.NOFA

NTT-CSL.m.balanced.1 : video.NOFA + audio.BALANCED

NTT-CSL.m.balanced.2 : video.BALANCED + audio.NOFA

NTT-CSL.m.balanced.3 : video.BALANCED + audio.BALANCED

5.3 Evaluation results

Figure 4 shows the evaluation results for the submitted runs. We found no significant difference between the the three runs for the balanced profile. As regards the result for the no false alarm profile, the actual NDCR for some transformation were very high (bad).

We examined the queries judged as false alarms, and found that the reference database contains some duplicate audio content, and this causes a false alarm when the video result is null, even if the audio result is a true positive.

6 Conclusions

In this paper, we described our approaches and results in relation to the TRECVID 2011 CBCD task. We introduced a very short window and time consistency filtering to cope with the detection of short segments. In our current system, the audio and video are processed independently. Our future tasks will include the fusion of the audio and video searches.

References

[1] T. Kurozumi, H. Nagano, and K. Kashino, "A robust video search method for video signal queries captured in the real world," *IEICE Trans. Inf.& Syst.(Japanese Edition)*, vol. J90-D, no. 8, pp. 2223–2231, 2007, in Japanese.

[2] K. Kashino, A. Kimura, H. Nagano, and T. Kurozumi, "Robust search methods for music signals based on simple representation," in *ICASSP 2007: Proceedings of 2007 International Conference on Acoustics, Speech, and Signal Processing*, 2007, pp. IV–1421–1424.

[3] R. Mukai, T. Kurozumi, K. Hiramatsu, T. Kawanishi, H. Nagano, and K. Kashino, "NTT Communication Science Laboratories at TRECVID 2010 content-based copy detection," in *TRECVID 2010 notebook papers*, 2010, pp. 340–349, http://www-nlpir.nist.gov/projects/tvpubs/tv10.papers/ntt-csl.pdf.

[4] H. Nagano, K. Kashino, and H. Murase, "A fast search algorithm for background music signals based on the search for numerous small signal components," in *ICASSP 2003: Proceedings of 2003 International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. V–796–799.

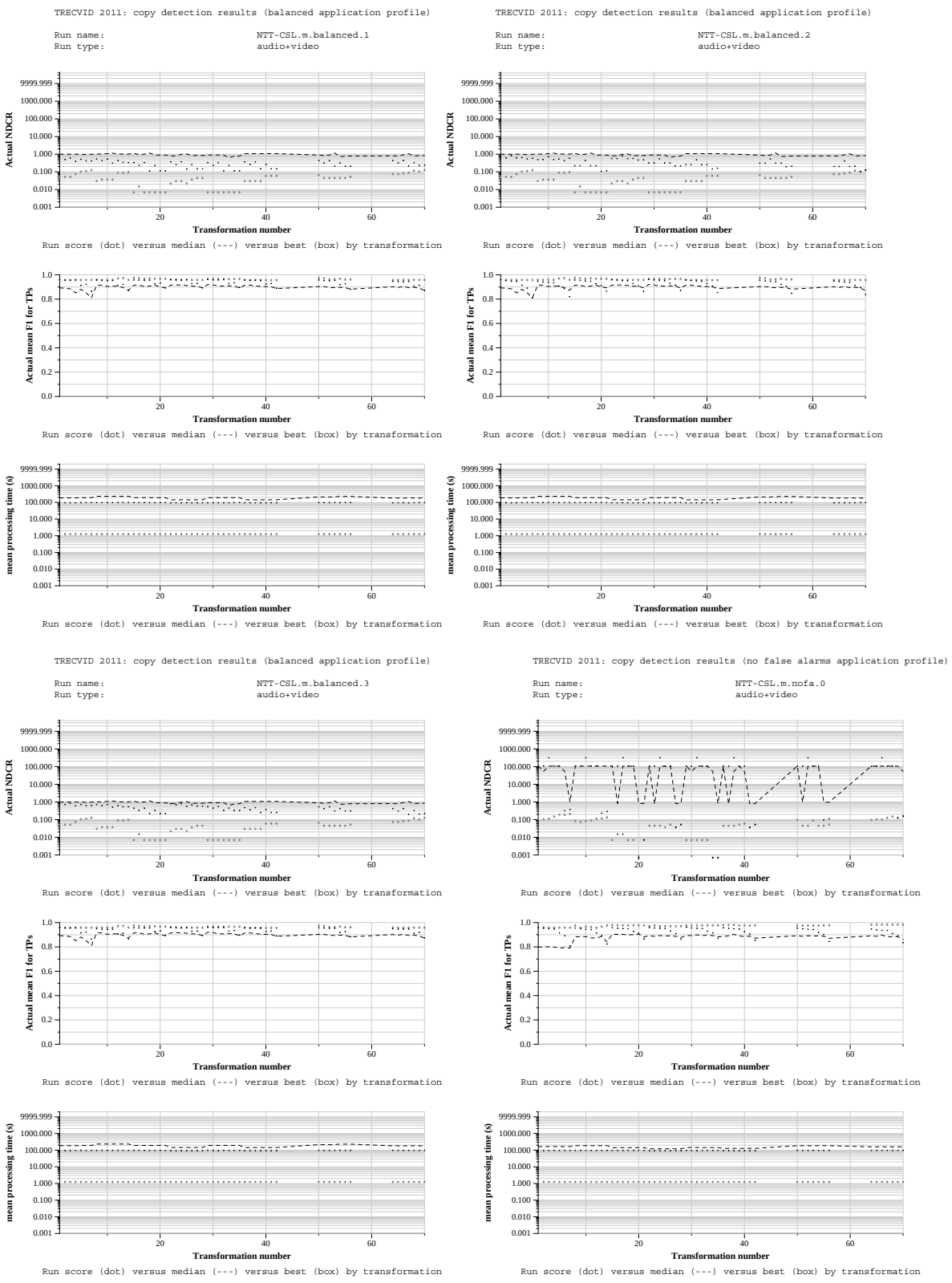

Figure 4: Evaluation results