TNO at TRECVID2008

Combining Audio and Video Fingerprinting for Robust Copy Detection

Peter Jan Doets, Pieter Eendebak, Elena Ranguelova, Wessel Kraaij
{Peter_Jan.Doets, Pieter.Eendebak, Wessel.Kraaij}@TNO.nl, E.Ranguelova@primevision.com
Netherlands Organisation for Applied Scientific Research (TNO)

Abstract

TNO has evaluated a baseline audio and a video fingerprinting system based on robust hashing for the TRECVID 2008 copy detection task. We participated in the audio, the video and the combined audio-video copy detection task. The audio fingerprinting implementation clearly outperformed the video fingerprinting implementation. We combined the audio fingerprinting results with the video fingerprint result, both from the TNO run, and from the submitted video run with the strongest correlation in the results.

1. Philips robust hash copy detection

The Philips Robust Hash (PRH) generates fingerprints consisting of a binary time-series [1], [2]. The PRH for audio and video use different features, but result in fingerprints with a comparable structure. Therefore the matching and database procedures are identical, although parameters may vary. The features are computed for each audio or video frame, and such a per-frame fingerprint is called a sub-fingerprint. A sequence of sub-fingerprints used for identification is called a fingerprint block. The hamming distance or Bit Error Rate (BER) is used to compute the distance between two fingerprint blocks.

In the enrollment (ingest) phase, a look-up table of sub-fingerprints is created for each clip. In the identification phase, (a selection of) the sub-fingerprints from the fingerprint block of the query fingerprint are matched against the database of pre-computed fingerprints. For more details, please refer to [1] and [2].

2. Video copy detection

For the video copy detection task we generated sub-fingerprints using the Haar based fingerprint described in [2]. This sub-fingerprint uses differences of average intensities in blocks in the images, and hence is robust to a number of video transformations (noise, affine intensity changes). A difference between subsequent frames is taken to into video motion into the fingerprint. From the sub-fingerprints a 16-bit index is generated for each of the video files.

A single sub-fingerprint does not have enough distinctive power to perform a search. Therefore for a given query video the index is used to find initial matches between the query and the database video. These initial matches contain some false matches and (hopefully) all the correct matches.

For each initial match we determine the matching score for an interval of fixed length between the query and database video using the sum of absolute differences between the sub-fingerprints. If the difference is below a threshold the interval is expanded. In this way the length of the match can grow untill the query and database video do not match any more. A final score of the matched interval is calculated using the BER and length of the interval. Finally the matched intervals are thresholded and combined into the final matches.

After receiving the TRECVID copy detection results, we discovered an error in the script generating the text document submitted to NIST. The error printed an incorrect query ID. Therefore, almost none of the results matched the ground truth tables used in the evaluation. Table 1 and Table 2 contain the results that have been corrected for this specific error. The tables have been generated by the evaluation software provided by NIST after the closing date of the copy detection task. The difference between TNO.v.1 and TNO.v.2. is the value of the threshold used to determine whether two intervals match or not.

From literature and analysis we know that the algorithm has difficulties when the reference material is geometrically or temporally distorted. This potentially includes transformation types 1 (camcording), 2 (PIP) and the mixture transform types 6 and 7 (including frame drops, ratio change), 8 and 9 (including crop, shift, flip) and 10 (5 randomly selected distortion). In the tables we can indeed see worst performance for transform types 1-2, and the composite distortions 7-10.

Literature suggests to use 32 bits/frame; in our implementation we used 16 bits/frame. Further analysis should show whether this implementation choice limited the discrimination capabilities of the fingerprints.

3. Audio copy detection

We used a baseline implementation of the audio PRH as described in [1]. Contrary to video, the audio has to be framed for processing by the algorithm. Every 11.6 msec a 32 bit sub-fingerprint is extracted based on two subsequent audio frames. The audio frames have strong overlap (approx 96%). The 32 bit sub-fingerprint is based on 33 log-spaced frequency bands in the range 300-2000 Hz. Each frequency band matches the base frequency of a tone. Each individual fingerprint block of 256 sub-fingerprints should be below a detection threshold. The confidence is computed over the entire detected fingerprint block.

In the TRECVID scenario, two factors contribute to the confidence: BER and interval length. We have chosen to relate the combined confidence score to the probability that the detection is actually a false positive. Therefore, we model the BER of two non-related fragments as a random number drawn from a normal distribution with average value 1/2 and the variance which is linear with the length of the fragments that are compared. We relate the confidence score to the probability density level of the before mentioned BER distribution.

Table 1 Detailed results for run TNO.v.1, bug-fixed after official submission

TNO.v.1.fixed	Transformations									
	1	2	3	4	5	6	7	8	9	10
Total_Queries	441	284	296	313	330	244	141	333	352	269
Mean_F1	0.619	0.000	0.720	0.594	0.731	0.603	0.391	0.818	0.000	0.119
Mean_proc_time	22948.89	20701.51	22460.93	23267.31	24130.68	21809.90	20983.85	23207.36	23758.94	23059.55
Total_proc_time	4612727	4161003	4514646	4676729	4850267	4383790	4217753	4664680	4775547	4634970
TP_count	2	0	17	20	28	33	21	4	0	1
Miss_count	132	134	117	114	106	101	113	130	134	133
FA_count	439	284	279	293	302	211	120	329	352	268
Min_NDCR	1.245	1.267	1.161	1.081	1.228	1.056	0.997	1.053	1.563	1.375

Table 2 Detailed results for run TNO.v.2, bug-fixed after official submission

TNO.v.2.fixed	Transformations									
	1	2	3	4	5	6	7	8	9	10
Total_Queries	217	94	164	108	223	60	16	136	146	66
Mean_F1	0.065	0.000	0.651	0.372	0.714	0.339	0.324	0.327	0.000	0.000
Mean_proc_time	6465.35	6131.74	7367.05	6515.33	8060.23	7067.80	5994.73	5945.62	6024.74	7045.96
Total_proc_time	1299536	1232480	1480777	1309582	1620107	1420627	1204941	1195069	1210972	1416237
TP_count	1	0	39	15	85	13	3	4	0	0
Miss_count	133	134	95	119	49	121	131	130	134	134
FA_count	216	94	125	93	138	47	13	132	146	66
Min_NDCR	1.315	1.064	0.897	0.993	0.644	0.993	1.000	1.120	1.131	1.071

By taking the log, and discarding some of the constant terms, we get the following confidence score:

```
Conf = log10(-(BER - 1/2)/ sqrt((const/len)))
```

Where 0 <= BER <= 1 and len is the fingerprint block length in audio frames, and the constant is the scaling factor between the variance and the length.

Table 3 shows the audio fingerprinting results.

Table 3 Detailed results for run TNO.a.1

	Transformations						
	1	2	3	4	5	6	7
Total_Queries	133	133	133	133	129	129	127
Mean_F1	0.957	0.961	0.963	0.962	0.688	0.681	0.605
Mean_proc_time	213.83	211.88	200.78	200.93	249.60	232.86	236.58
Total_proc_time	42980	42587	40356	40387	50169	46804	47553
TP_count	131	131	131	131	127	127	125
Miss_count	3	3	3	3	7	7	9
FA_count	2	2	2	2	2	2	2
Min_NDCR	0.019	0.019	0.019	0.019	0.049	0.049	0.064

Analysis shows that:
- Actually one query did not result in a detected copy for all transformation types;
- Two queries resulted in the same wrong identification result for all transformations;
- Four queries were made undetectable by transformations 5-7;
- Transformation 7 rendered two additional queries undetectable.

Transformations 1-4 mainly encompass compression (mp2, mp3), companding, bandpass filtering. Transformations 5-7 all include mixture with speech, with unknown mixing weights.

4. Combined audio-video detection

Audio and video fingerprinting systems can be combined at different levels, e.g. at the feature level, the fingerprint level or at the decision level. Due to the nature of our results and the combination with a run which was not our own submission, we chose a combination at decision level. For the first two combination methods, the audio and video fingerprinting systems should also be temporally aligned.

Analysis of the submitted audio and video runs showed little correlation between the runs. Therefore, we submitted both a combination of our video and audio run, as well as a combination of our audio run, and the video run which showed most overlap with our audio run.

Since the combination is based on the individual audio and video copy detector outputs, two choices determine the effect of the combination:

a. the operator used to combine the results
b. how to deal with conflicting audio and video results (can be a result of step a)
c. how to compute the joint confidence score

The individual confidence scores are normalized as follows. A joint-Gaussian PDF for the confidence scores is assumed. Since we did not measure a strong correlation between the audio and video confidence scores, we chose to normalize individually such that each confidence score have average equal to 0 and variance equal to 1. The joint confidence is computed as the summation of the individual normalized scores. We then added some positive constant to make sure that the resulting confidence score is positive. A remark should be made about the fact that the only the confidence scores are known that have resulted in a hypothesized positive detection.

The confidence of the individual copy detection result corresponds to the entire detection copy interval. However, in the combination only a part of this may be considered. Without any knowledge of the underlying confidence measure, it becomes difficult to estimate the partial confidence for part of the interval. Furthermore, the relation between interval length and confidence score may be non-linear, as in the confidence score used in the audio copy detection run.

TNO has submitted the following combination runs:

TNO.m.A01   intersection of TNO.a.01 and TNO.v.01
TNO.m.A02   'or' of TNO.a.01 and TNO.v.01, with removal of overlap
TNO.m.B01   intersection of TNO.a.01 and INRIA-LEAR.v.Soft.

Since the submitted TNO.v.1. run contains almost no correct results due to a processing error, we concentrate the result analysis on the TNO.m.B01 run. This run is based on the INRIA-LEAR.v.Soft video run and the TNO.a.01 audio run.

In the experiments, we used an intersection operation. That is, the audio and video results should point to the same reference video, and only the overlap is considered to be correct. In this way, the confidence is raised, and the false positive detection rate is lowered or kept constant. The number of false misses, however, potentially increases. For instance, if either the audio or video is correct, but not both, the result is still discarded, irrespective of the associated confidence scores.

This is illustrated by comparing the audio run with the audio-video run, averaged over the video transformations. Both are shown in Table 4 and Table 5.

Table 4 Selection of results for run TNO.a.1

TNO.a.1	Transformation						
	1	2	3	4	5	6	7
Total_Queries	133	133	133	133	129	129	127
TP_count	131	131	131	131	127	127	125
Miss_count	3	3	3	3	7	7	9
FA_count	2	2	2	2	2	2	2
Min_NDCR	0.019	0.019	0.019	0.019	0.049	0.049	0.064

Table 5 Selection of results for run TNO.m.B01, averaged over the video transformation types

TNO.m.B01	Transformation						
--> audio view	1	2	3	4	5	6	7
Total_Queries	127.4	128.1	128.6	128.1	125.4	129.7	120.1
TP_count	126.4	127.1	127.6	127.1	123.7	128.7	119.1
Miss_count	7.6	6.9	6.4	6.9	10.3	5.3	14.9
FA_count	1	1	1	1	1.7	1	1
Min_NDCR	0.0498	0.0443	0.0405	0.0443	0.0651	0.0323	0.1042

Table 6 Selection of results for run INRIA.LEAR.v.Soft.results, averaged over the video transformation types

INRIA-LEAR.v.Soft.results	Transformation									
	1	2	3	4	5	6	7	8	9	10
Total_Queries	794	338	336	197	266	326	406	613	668	665
TP_count	131	130	134	132	134	130	129	134	132	122
Miss_count	3	4	0	2	0	4	5	0	2	12
FA_count	663	208	202	65	132	196	277	479	536	543
Min_NDCR	0.126	0.046	0.015	0.038	0.012	0.069	0.115	0.045	0.080	0.246

Table 7 Selection of results for run TNO.m.B01, averaged over the video transformation types

TNO.m.B01	Transformation									
--> video view	1	2	3	4	5	6	7	8	9	10
Total_Queries	128.1	127.9	127.9	126.3	127	126.4	126.1	125.6	125.7	126.7
TP_count	127.1	126.9	126.7	125.1	125.9	125.3	125	124.4	124.6	125.7
Miss_count	6.9	7.1	7.3	8.9	8.1	8.7	9	9.6	9.4	8.3
FA_count	1	1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1
Min_NDCR	0.0440	0.0461	0.0463	0.0581	0.0527	0.0570	0.0590	0.0633	0.0623	0.0547

The same holds for the comparison of the INRIA video run with the audio-video run, averaged over the audio transformations shown in Table 7. The video run table was generated using the NIST evaluation software on the INRIA run submission.

We hypothesize that the intersection operation for combining audio and video results at decision level is probably a too crude and conservative operation, further investigation is needed for finding more effective combination methods.

5. Discussion

Fingerprint size/rate is an essential parameter which should be taken into account when comparing fingerprinting systems. Within the capabilities of a certain approach (e.g. robustness to certain distortions, etc.), the fingerprint size usually determines the performance vs. search time trade-off. A larger fingerprint can lower the false positive (and false negative) probability of a detection, but increases the search space since more unique fingerprints can be represented by a larger fingerprint.

Although audio fingerprinting seems to be more mature and the complexity of distortions might be lower, it is hard to say that ‘audio fingerprinting is easier than video fingerprinting’. In the TRECVID copy detection task, the distortions that were considered for audio were not that severe. For instance, there were no distortions that effectively scale time of frequency. For video, the equivalent would be distortions that change the geometrical of temporal structure of the video. Examples of these kinds of distortions are horizontal flip, picture-in-picture, camcording and frame drops. Some of these distortions actually preserve the geometrical structure of the reference material, but the overall geometrical structure of the query is different.

Acknowledgements

This research was funded by the EU safer internet project I-dash (the Investigator's dashboard), which deals with the detection of child abuse video in seized video collections.

References

[1] J. Haitsma and T. Kalker, “A Highly Robust Audio Fingerprinting System”, *3rd International Conference on Music Information Retrieval (ISMIR),* October 2002.

[2] J. Oostveen, T. Kalker and J. Haitsma, “Feature Extraction and a Database Strategy for Video Fingerprinting”, *5th International Conference VISUAL*, LNCS, vol. 2314, pp. 117 – 128, October 2002.