

# Evaluating Detection of Near Duplicate Video Segments

Werner Bailer  
Institute of Information Systems  
JOANNEUM RESEARCH Forschungsgesellschaft mbH  
Steyrergasse 17, 8010 Graz, Austria  
werner.bailer@joanneum.at

## ABSTRACT

The automatic detection of near duplicate video segments, such as multiple takes of a scene or different news video clips showing the same event, has received growing research interest in recent years. However, there is no agreed way of evaluating near duplicate detection algorithms. This makes it very hard to compare the performance of different algorithms, even if they are applied to the same data set. In this paper we have implemented several evaluation measures found in literature and we apply them to real algorithm outputs and a simulated result data set. We then calculate the correlation between the results obtained with the different measures in order to investigate whether they can be compared or not. The results show that the correlation between the measures is some cases quite low, and some measures are especially sensitive to certain types of deviations from the ground truth. However, a group of precision/recall type measures and two others are clearly correlated, though with moderate correlation coefficients. We also analyze the correlation between these measures and the subjective human judgment of the number of repeated segments in summary videos.

## Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval; I.5.3 [Computing Methodologies]: Pattern Recognition—*Similarity measures*

## General Terms

Experimentation, measurement, performance

## Keywords

Scene clustering, copy detection, repeated content

## 1. INTRODUCTION

The automatic detection of near duplicate video segments, such as multiple takes of a scene or different news video clips

showing the same event, has received growing research interest in recent years. The problem of near duplicate detection differs from typical video copy detection problems (e.g. detection of pirated copies of video content) as it does not aim at detecting a (possibly severely distorted) copy of the same content, but a piece of video with similar but not identical content. For example, in news video, the same action may be shot from a slightly different viewpoint. In different takes of a scene of the actors might move differently, or the scene may have a different timing. The problem of distortions is in most applications less critical than for classical copy detection applications, but in both applications the detection of partial matches is important. In many cases, near duplicate detection also involves clustering matches so that a cluster contains segments that are mutually near duplicates.

When looking into the literature on near duplicate video detection it becomes apparent that there is no agreed way of evaluating near duplicate detection algorithms. Some researchers use different variants of precision/recall, others use measures inspired by information theory and some benchmarks use measures based on task dependent costs. This makes it very hard to compare the performance of different algorithms, even if they are applied to the same data set. Some researchers attempt to compare results using for example different types of precision/recall measures (e.g. [6]), and it is not clear whether this is justified.

In this paper we have implemented several evaluation measures found in literature and we apply them to different data sets. We then calculate the correlation between the results obtained with the different measures in order to investigate whether they can be compared or not. In addition, we analyze the correlation between these measures and the subjective judgment of the number of repeated segments in summary videos.

The rest of this paper is organized as follows. Section 2 reviews the different measures for evaluating near duplicate detection found in literature. Section 3 describes the experiments we perform to compare the measures, presents the results and discusses them. Section 4 concludes the paper.

## 2. REVIEW OF MEASURES

In this section we review three variants of precision/recall based measures, a measure based on normalized mutual information, two measures used in the MUSCLE VCD benchmark [8] and the one used in the TRECVID benchmark [15] content-based copy detection task. The latter is the only one of those that has to our knowledge not been applied to near duplicate detection, but is included for comparison due

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIVR '10, July 5-7, Xi'an China

Copyright ©2010 ACM 978-1-4503-0117-6/10/07 ...\$10.00.

**Table 1: Comparison of some properties of the different measures: supports frame precise measurement (frame), supports evaluation of cluster structure of near duplicates (cluster), and does not require the same segmentation for ground truth and result (segment).**

	frame	cluster	segment
PR-A		x	x
PR-S			
PR-F	x	x	x
NMI		x	
M-S			
M-F	x		x
NDCR			

to the popularity of the TRECVID benchmark. There are further methods used for evaluating near duplicate detection that require additional metadata (e.g. topic labels), such as the measure for evaluating news story threading in [7]. Thus they are not generally applicable and have not been considered in this paper.

In this paper we use the following notation. We define the set  $S_V = \{s_1, \dots, s_N\}$  of segments (shots, subshots) of a video or video collection  $V$ . The aim of near duplicate detection is to identify segments  $d$  that are near duplicates of other segments in  $S_V$ , where depending on the algorithm a duplicate segment  $d_i$  may have the same boundaries as an input segment  $s_i$  or not. In addition, many algorithms will group duplicate segments into  $M$  clusters  $D_j = \{d_1, \dots, d_K\}$ , so that the segments in  $D_j$  are near duplicates of all other segments in  $D_j$ , but not near duplicates of any segment not in  $D_j$ . Such a cluster corresponds e.g. to a scene or a news topic thread. We denote the set of all such clusters as  $\mathcal{D} = \{D_1, \dots, D_M\}$ . We denote the ground truth clustering as  $\mathcal{D}'$ , and similarly all other prime decorated identifiers refer to the respective ground truth variables.

Throughout the paper the following abbreviations are used for the different measures: precision/recall of aligned clusters (PR-A), precision/recall on shot/segment basis (PR-S), precision/recall on frame basis (PR-F), normalized mutual information (NMI), Muscle VCD segment metric (M-S), Muscle VCD frame metric (M-F) and TRECVID CBCD normalized detection cost rate (NDCR). In the following we discuss each of the measures in detail. Table 1 summarizes some properties of the different measures.

## 2.1 Precision and Recall Based Measures

### 2.1.1 Aligned Clusters

This method has been proposed in [2] for evaluating detection and clustering of repeated takes of a scene. The number of takes correctly and falsely assigned to one scene (including partial takes) is counted. The exact temporal extent of the identified takes as well as the alignment of partial takes is not taken into account. The authors argue that this would enormously increase the effort for creating the ground truth and that the correct exact temporal extent is difficult to define even for a human in many cases.

As the order of scenes in  $\mathcal{D}$  and  $\mathcal{D}'$  is not known, the first step is to align the scenes. This is done by assigning

each result scene  $D$  to that scene  $D'$  of the ground truth  $\mathcal{D}'$ , for which the overlap between the takes of the clusters is maximized:

$$\text{matching\_scene}(D, \mathcal{D}') = \operatorname{argmax}_{D' \in \mathcal{D}'} \sum_{i=1}^M \sum_{j=1}^{M'} |d_i \cap d'_j| \quad (1)$$

where  $|d_i \cap d'_j|$  calculates the temporal overlap between the takes in frames. Additionally, a scene from the result set  $D$  may only be assigned to at most one ground truth scene  $D'$  and vice versa. After this step there can be unassigned scenes, i.e. `matching_scene` returns an empty set.

Once the clusters have been assigned, the number of temporally overlapping takes is counted, i.e. we determine

$$n_{\text{overlap}} = \sum_{D \in \mathcal{D}} |D \cap \text{matching\_scene}(D, \mathcal{D}')| \quad (2)$$

If a take of the ground truth overlaps with more than one take of the same result cluster, only a single overlap is counted. The takes of an unassigned scene  $D$  are counted as false positives, the takes of an unassigned scene  $D'$  are counted as false negatives. From the sum of correct overlaps of all clusters, precision and recall are calculated as

$$\text{precision} = \frac{n_{\text{overlap}}}{\sum_{D \in \mathcal{D}} |D|}, \text{ and} \quad (3)$$

$$\text{recall} = \frac{n_{\text{overlap}}}{\sum_{D' \in \mathcal{D}'} |D'|}. \quad (4)$$

### 2.1.2 Shot/Segment Basis

The authors of [5] propose a method for matching clips in video databases using color similarity and edit distance. The authors evaluate the approach by detecting near duplicate shots in a news video database and detecting near duplicate advertisement clips. The performance measure is the precision/recall rate of correctly identified near duplicate shots. The measure is based on the assumption that the segment boundaries in the ground truth and in the result are identical. The same measure is used in [14] and for a near duplicate discovery experiment in [9].

The algorithm being evaluated does not cluster matching segments, but yields for each segment  $s_i$  a result set  $R_i$  of near duplicate segments. Thus each segment  $s'_i = s_i$  in a ground truth cluster  $D'_j$  is expected to produce a set of  $|D'_j| - 1$  correct results, i.e.  $|D'_j|(|D'_j| - 1)$  results in total for the segments in cluster  $D'_j$ . Precision and recall are then defined as

$$\text{precision} = \frac{\sum_{j=1}^{M'} \sum_{s_i \in D'_j} |(D'_j \setminus s_i) \cap R_i|}{\sum_{i=1}^N |R_i|}, \text{ and} \quad (5)$$

$$\text{recall} = \frac{\sum_{j=1}^{M'} \sum_{s_i \in D'_j} |(D'_j \setminus s_i) \cap R_i|}{\sum_{j=1}^{M'} |D'_j|(|D'_j| - 1)}. \quad (6)$$

### 2.1.3 Frame Basis

The authors of [6] propose a method for scene clustering based on sequence matching. They evaluate their method by calculating the per frame precision and recall based on

the scene assignment of frames. As the calculation is only based on assignment of single frames, no assumptions about the temporal segmentation of the result are necessary. It is only assumed that the corresponding scenes  $D_j$  and  $D'_j$  are known (otherwise the alignment could be determined similar as in Section 2.1.1). Precision and recall are then defined as

$$\text{precision} = \frac{\sum_{j=1}^{M'} \sum_{d'_k \in D'_j} \sum_{d_l \in D_j} |d'_k \cap d_l|}{\sum_{j=1}^{M'} \sum_{d_k \in D_j} |d_k|}, \text{ and} \quad (7)$$

$$\text{recall} = \frac{\sum_{j=1}^{M'} \sum_{d'_k \in D'_j} \sum_{d_l \in D_j} |d'_k \cap d_l|}{\sum_{j=1}^{M'} \sum_{d'_k \in D'_j} |d'_k|}, \quad (8)$$

where  $|d_k|$  denotes the length segment  $d_k$  in frames.

## 2.2 Normalized Mutual Information (NMI)

An approach for clustering repeated takes into scenes is reported in [13]. The authors propose an evaluation measure based on normalized mutual information (NMI) [4]. The scene clusters of the result  $\mathcal{D}$  and the ground truth  $\mathcal{D}'$  are interpreted as random variables. In case of maximum likelihood estimation, their mutual information is given as

$$I(\mathcal{D}; \mathcal{D}') = \sum_{i=1}^M \sum_{j=1}^{M'} \frac{|D_i \cap D'_j|}{N} \log \left( \frac{N|D_i \cap D'_j|}{|D_i||D'_j|} \right), \quad (9)$$

where  $|D_i \cap D'_j|$  is the number of segments shared between the scenes, assuming that the temporal segmentations of the ground truth and the result are identical. The mutual information is normalized to obtain NMI:

$$NMI(\mathcal{D}, \mathcal{D}') = 2 \frac{I(\mathcal{D}; \mathcal{D}')}{H(\mathcal{D}) + H(\mathcal{D}')}, \quad (10)$$

with  $H(\cdot)$  being the entropy of a cluster:

$$H(\mathcal{D}) = - \sum_{i=1}^N \frac{|D_i|}{N} \log \frac{|D_i|}{N}. \quad (11)$$

## 2.3 Muscle VCD Metrics

The video copy detection approaches described in [12, 16] use the metric defined by the Muscle VCD benchmark for locating segments in a stream. The methods do not produce clustering of results, but yield for each segment  $s'_i = s_i$  in a ground truth cluster  $D'_j$  a result set  $R_i$  of matching segments. Two quality measures are defined in [8]. A segment based measure is defined as

$$q_s = \frac{N_{correct} - N_{falseAlarm}}{N}, \quad (12)$$

where

$$N_{correct} = \sum_{j=1}^{M'} \sum_{s_i \in D'_j} |(D'_j \setminus s_i) \cap R_i| \quad (13)$$

and

$$N_{falseAlarm} = \sum_{j=1}^{M'} \sum_{s_i \in D'_j} |R_i| - |(D'_j \setminus s_i) \cap R_i|. \quad (14)$$

The second measure is frame based and defined as one minus the fraction of mismatched (missed, imprecisions, false positives) and all frames in the queries:

$$q_f = 1 - \frac{N_{mis}}{\sum_{i=1}^N |s_i|}, \quad (15)$$

where

$$N_{mis} = \sum_{j=1}^{M'} \sum_{s_i \in D'_j} \left( \sum_{d'_k \in (D'_j \setminus s_i)} |d'_k| + \sum_{d_k \in R_i} |d_k| - 2 \sum_{d_k \in R_i, d'_l \in (D'_j \setminus s_i)} |d_k \cap d'_l| \right). \quad (16)$$

## 2.4 Normalized Detection Cost Rate (NDCR)

NDCR is used in the TRECVID [15] content based copy detection task. The measure involves a target false alarm rate  $R_{Target}$  per query duration and costs for false alarms  $C_{FA}$  and misses  $C_{Miss}$ , which are combined into a parameter  $\beta = C_{FA}/(C_{Miss}R_{Target})$ . The costs and target false alarm rate depend on the application. Here we use the balanced profile defined in the TRECVID 2009 CBCD task, which assigns equal costs to false alarms and misses:  $C_{FA} = C_{Miss} = 1$ ,  $R_{Target} = 0.5/\text{hour}$ . Algorithms are assumed to yield for each segment  $s'_i = s_i$  in a ground truth cluster  $D'_j$  a result set  $R_i$  of matching segments. The measure is then defined as

$$NDCR = P_{Miss} + \beta R_{FA}, \quad (17)$$

where the probability of a miss is defined as

$$P_{Miss} = \frac{\sum_{j=1}^{M'} \sum_{s_i \in D'_j} |D'_j| - |(D'_j \setminus s_i) \cap R_i|}{\sum_{j=1}^{M'} |D'_j| (|D'_j| - 1)}. \quad (18)$$

The hourly false alarm rate is defined as

$$R_{FA} = \frac{\sum_{j=1}^{M'} \sum_{s_i \in D'_j} |R_i| - |(D'_j \setminus s_i) \cap R_i|}{\sum_{j=1}^{M'} \sum_{d'_k \in D'_j} \frac{|d'_k|}{3600f}}, \quad (19)$$

with  $f$  being the frame rate of the query videos.

## 3. EVALUATING THE CORRELATION BETWEEN MEASURES

In this section we describe the experiments we have performed to compare the evaluation measures and discuss the results. We use data from the TRECVID BBC rushes 2007 task [11], which consists of 42 videos containing rushes from different BBC productions.

## 3.1 Experiments

The first two experiments measure the correlation among the measures on the outputs of a real and a simulated data set. The third experiment uses a set of summaries that has been generated for the videos in the test set, and analyzes the correlation between the measures and a rating of the amount of repeated content given by a human evaluator. For precision/recall type measures we calculate  $F1 = \frac{2pr}{p+r}$  from precision  $p$  and recall  $r$ , which can then be used for calculating the correlations with measures that yield just a single value.

### 3.1.1 Correlation on Repeated Take Detection Algorithm Outputs

This experiment uses the results of the algorithms described in [3]. The algorithms are two variants of the longest common subsequence (LCSS) based algorithm presented in that paper, and an algorithm based on dynamic time warping (DTW) used for comparison. All results are based on a ground truth temporal segmentation in order to support algorithms that need same the segmentation of ground truth and result set. The data consists of the outputs of the three algorithms on six videos from the TRECVID 2007 rushes test set (MRS07063, MRS025913, MRS044731, MRS144760, MRS157475, MS216210) and the ground truth created by the authors of [3].

### 3.1.2 Correlation on Simulated Results

This experiment uses the complete TRECVID 2007 rushes test set and the ground truth provided by NHK [10]. The simulated result sets are generated by applying transformations to the ground truth: dropping, adding or shifting segments. Dropping and adding segment uses a parameter that specifies the fraction of segments to be dropped or added. The actual number  $k$  of segments to be dropped or added is determined by drawing from a normal distribution centered around the parameter value. For dropping, the segments in the ground truth are randomly permuted and the first  $k$  segments are deleted. For adding, a list of all possible segments not yet in the ground truth is created and the first  $k$  of the randomly permuted list of segments are added to either existing clusters or new clusters. The shift transformation uses as parameter the maximum number of frames  $f$  by which the segment boundaries are modified. For each segment, a random number is drawn from the uniform distribution in the interval  $[-f, f]$  and the segments are shifted by this value. Note that the transformations are on the *output segments* of a near duplicate detector and not on the videos, i.e. they are independent of the type of content transformation (e.g. brightness or color change, text overlay, framing, etc.) that might have caused the detector to report a different segment.

The following parameter values have been used: dropping 50% of the segments, adding 100% of the segments (i.e. dropping and adding each change the number of segments on average by factor 2) and shifting segment boundaries by  $\pm 30$  frames. Four sets of results have been generated: one for each of the single transformations and one where all transformations have been applied. For each video 100 results have been generated for each of the four sets, i.e. they contain 4200 results each.

### 3.1.3 Correlation with Human Perception

This experiment is performed on the summaries of the TRECVID 2007 rushes test set, created with the algorithm described in [1] according to the rules of the TRECVID evaluation. The data set contains 42 summaries. As part of the evaluation by NIST, the human evaluators rated the number of duplicate segments on a scale from 0 (many near duplicate segments) to 5 (no near duplicate segments). The values in the test set are in the range 2.33 to 5.00, with a mean of 3.78, median of 3.67 and standard deviation of 0.75.

The algorithm creating the summaries assumes that no duplicates were included, thus the duplicate cluster structure describing the summary puts every segment in an own cluster. We have created ground truth, identifying the repeated segments still found in the summaries. The duplicate cluster description of the summaries is matched against this ground truth using the different evaluation measures and their correlation with the human evaluator is then analyzed.

## 3.2 Results

The tables in this section present the correlation coefficients between the different types of evaluation measures. For most of the results we present both Pearson's product moment correlation<sup>1</sup> coefficient and Spearman's rank correlation coefficient<sup>2</sup>, as the ranking might still be comparable, even if no linear correlation between measures is given.

Table 2 shows the correlations among the F1 values of the precision/recall measures and the outputs of the other measures on the results from the repeated take detection algorithm. Table 3 shows the correlations among the precision and recall values of the precision/recall measures and the outputs of the other measures.

Table 4 shows the correlations among the F1 values of the precision/recall measures and the outputs of the other measures on the simulated result data set. The correlations are given separately for result sets generated by applying only one of the three transformations and a result set generated by applying all the transformations. Table 5 shows the correlations among the precision and recall values of the precision/recall measures and the outputs of the other measures.

Figure 1 shows a plot of the (F1) measures against the human evaluator judgments of the amount of repeated content in the summary videos. Table 6 shows the correlations of the F1 values of the precision/recall measures and the outputs of the other measures with the human evaluator judgment on this set of summaries. Note that as every incorrectly associated segment is counted both as a false positive and a false negative,  $F1 = p = r$ . For the PR-S measure, the result sets for querying each of the segments in the summary are empty, as each segment is assumed to be the only representative of a set of duplicates. Consequently the measure always yields  $p = r = 0$  in these cases and has thus not been included in this experiment.

<sup>1</sup>Pearson's correlation coefficient between  $N$  samples from two variables  $X$  and  $Y$  is defined as the covariance of the two variables divided by the product of their standard deviations:  $r_{X,Y} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}$ .

<sup>2</sup>Spearman's rank correlation coefficient is defined as Pearson's product moment correlation coefficient between the ranks of two variables.

**Table 2: Correlation of (F1 values of) measures on the repeated take detection algorithm results. The upper right of the table contains Pearson’s product moment correlation coefficients, the lower right Spearman’s rank correlation coefficients. Coefficients with a confidence interval  $p < 0.10$  are written bold.**

	PR-A	PR-S	PR-F	NMI	M-S	M-F	NDCR
PR-A		0.336	<b>0.920</b>	0.397	<b>0.999</b>	-0.000	<b>-0.790</b>
PR-S	0.321		<b>0.583</b>	<b>0.869</b>	0.308	<b>0.646</b>	-0.268
PR-F	<b>0.878</b>	0.561		0.580	<b>0.914</b>	0.315	<b>-0.856</b>
NMI	0.424	<b>0.930</b>	<b>0.638</b>		0.373	0.555	-0.209
M-S	<b>0.992</b>	0.271	<b>0.870</b>	0.400		-0.014	<b>-0.800</b>
M-F	0.037	<b>0.621</b>	0.391	<b>0.526</b>	-0.147		-0.273
NDCR	<b>-0.882</b>	-0.375	<b>-0.899</b>	<b>-0.410</b>	<b>-0.878</b>	-0.317	

**Table 3: Correlation of precision (upper table) and recall (lower table) of measures on the repeated take detection algorithm results (Pearson’s product moment correlation coefficients). Coefficients with a confidence interval  $p < 0.10$  are written bold.**

	PR-S	PR-F	NMI	M-S	M-F	NDCR
<i>Precision</i>						
PR-A	<b>0.629</b>	<b>0.927</b>	<b>0.332</b>	<b>0.998</b>	-0.035	<b>-0.806</b>
PR-S		0.492	<b>0.754</b>	<b>0.669</b>	<b>0.411</b>	<b>-0.543</b>
PR-F			0.140	<b>0.909</b>	0.002	<b>-0.889</b>
<i>Recall</i>						
PR-A	<b>-0.468</b>	<b>0.811</b>	<b>0.453</b>	<b>0.990</b>	0.031	<b>-0.768</b>
PR-S		0.052	<b>0.412</b>	-0.548	<b>0.599</b>	<b>0.404</b>
PR-F			<b>0.802</b>	<b>0.754</b>	<b>0.472</b>	<b>-0.676</b>

**Table 6: Correlation between (F1) measures and human evaluator judgments of the amount of repeated content in summary videos. The upper row contains Pearson’s product moment correlation coefficients ( $r$ ), the lower row Spearman’s rank correlation coefficients ( $\rho$ ). Coefficients with a confidence interval  $p < 0.10$  are written bold.**

	PR-A	PR-F	NMI	M-S	M-F	NDCR
$r$	0.211	<b>0.312</b>	0.250	0.211	0.219	-0.173
$\rho$	<b>0.284</b>	<b>0.370</b>	0.202	<b>0.284</b>	0.241	-0.167

### 3.3 Discussion

In all of the results, the negative correlations between NDCR and all the other measures are noticeable. The reason is that NDCR measures cost in contrast to the other measures, thus the negative correlation can be expected. All other negative correlations coefficients have small values, so that the respective measures can be assumed to be uncorrelated (the exception being some recall values, which are discussed below).

On the F1 measures of the repeated take detection results, the product moment and rank correlation coefficients yield similar results. Several measures are clearly correlated with coefficients over 0.9, while some are completely uncorrelated. There are no stronger correlations among the frame and segment based measures than between the two groups, however, another grouping emerges. The measures PR-A, PR-F, M-S and NDCR on the one side and PR-S, NMI and M-F on the other side form groups that are highly correlated among them and only weakly correlated between them. The overall picture of the correlations of precision and recall of the repeated take detection outputs is the same as for F1,

showing also the same grouping. For NMI, the precision is uncorrelated with PR-F, but has a strong correlation in terms of recall. The correlation of recall of PR-S with PR-A and M-S is negative, which is caused by results containing many duplicate segments. In this case the result segments overlap with many ground truth segments, yielding low precision and high recall in terms of PR-S. The other measures count only one match between a result and a ground truth segment in the entire result set, thus they also yield low recall in this case.

On the simulated results, rank correlation is in most cases higher than product moment calculation, and as in the first experiment, the correlation among segment and frame based measures is not generally higher than between them. However, the grouping based on the F1 measures changes. PR-S has a higher correlation with precision/recall measures than in the first experiment, probably due to the fact that the simulated results do not contain results with more than twice as many segments than the ground truth, while this occurs in the algorithm results. As in the first experiment, M-S is highly correlated with the precision/recall measures. In contrast to the first experiment, NDCR has a much lower correlation with the precision/recall measures and M-S, although these correlations are significantly higher (up to  $-1.000$ ) when only a single transformation is applied.

Dropping segments from the result is the transformation with the least impact on the measures. Most correlations are relatively high. The only exception is the M-F score, as the fraction of misclassified frames increases due to the lower size of the result set and the reduction of the correct results. M-F calculated as one minus the fraction of misclassified segments is then less correlated to the other measures. Another interesting result is that due to a number of smaller changes in ranking the rank correlation between M-S and

Table 4: Correlation of (F1 values of) measures on the simulated results created by applying all transformations (top) or each transformation individually (3 lower tables). The upper right of each table contains Pearson’s product moment correlation coefficients, the lower right Spearman’s rank correlation coefficients. Coefficients with a confidence interval  $p > 0.005$  are written italic.

<i>all</i>	<b>PR-A</b>	<b>PR-S</b>	<b>PR-F</b>	<b>NMI</b>	<b>M-S</b>	<b>M-F</b>	<b>NDCR</b>
<b>PR-A</b>		0.638	0.664	0.380	0.907	0.272	-0.243
<b>PR-S</b>	0.652		0.567	0.552	0.563	0.170	-0.216
<b>PR-F</b>	0.654	0.561		0.371	0.631	0.603	-0.376
<b>NMI</b>	0.471	0.618	0.425		0.340	<i>0.041</i>	-0.216
<b>M-S</b>	0.950	0.590	0.641	0.407		0.414	-0.479
<b>M-F</b>	0.304	0.255	0.595	0.103	0.431		-0.632
<b>NDCR</b>	-0.276	-0.228	-0.380	-0.192	-0.479	-0.675	
<i>drop</i>	<b>PR-A</b>	<b>PR-S</b>	<b>PR-F</b>	<b>NMI</b>	<b>M-S</b>	<b>M-F</b>	<b>NDCR</b>
<b>PR-A</b>		0.805	0.775	0.934	0.947	0.150	-0.989
<b>PR-S</b>	0.827		0.638	0.723	0.788	0.256	-0.805
<b>PR-F</b>	0.768	0.634		0.723	0.735	0.169	-0.768
<b>NMI</b>	0.956	0.756	0.736		0.903	0.249	-0.940
<b>M-S</b>	0.991	0.819	0.759	0.948		0.177	-0.973
<b>M-F</b>	0.151	0.268	0.169	0.250	0.158		-0.167
<b>NDCR</b>	-1.000	-0.827	-0.768	-0.956	-0.416	-0.151	
<i>add</i>	<b>PR-A</b>	<b>PR-S</b>	<b>PR-F</b>	<b>NMI</b>	<b>M-S</b>	<b>M-F</b>	<b>NDCR</b>
<b>PR-A</b>		0.620	0.570	0.543	0.933	0.506	-0.750
<b>PR-S</b>	0.910		0.285	0.309	0.608	0.248	-0.479
<b>PR-F</b>	0.766	0.684		<i>0.016</i>	0.533	0.936	-0.711
<b>NMI</b>	0.870	0.769	0.623		0.579	<i>0.014</i>	-0.212
<b>M-S</b>	1.000	0.910	0.766	0.870		0.483	-0.729
<b>M-F</b>	0.769	0.687	0.998	0.626	0.769		-0.657
<b>NDCR</b>	-0.896	-0.820	-0.825	-0.745	-0.896	-0.828	
<i>shift</i>	<b>PR-A</b>	<b>PR-S</b>	<b>PR-F</b>	<b>NMI</b>	<b>M-S</b>	<b>M-F</b>	<b>NDCR</b>
<b>PR-A</b>		0.557	0.266	0.531	1.000	0.676	-0.916
<b>PR-S</b>	0.527		0.674	0.816	0.557	0.474	-0.600
<b>PR-F</b>	0.217	0.681		0.673	0.266	0.515	-0.466
<b>NMI</b>	0.484	0.835	0.671		0.531	0.579	-0.611
<b>M-S</b>	1.000	0.527	0.217	0.484		0.676	-0.916
<b>M-F</b>	0.408	0.655	0.844	0.655	0.408		-0.728
<b>NDCR</b>	-0.976	-0.592	-0.341	-0.552	-0.976	-0.482	

NDCR drops significantly, while the product moment correlation is still quite high.

Adding segments has an overall slightly stronger impact on the correlations. The M-F measure is less affected by adding segments (at same size of ground truth and same number of correct segments). As NMI takes the size and entropy of both ground truth and result clusters into account, it is stronger affected by adding segment (and possibly clusters) to the result.

Shifting has the strongest impact of all the single transformations. One would expect that shifting (i) has a stronger impact on frame based measures than on segment based ones and (ii) impacts the correlations between frame and segment based measures. The first is not true, as the applied shift also influences the association of ground truth and result segments in some of the segment based measures and thus also decrease their scores. The second assumption holds, as we find the strongest correlation among segment based (e.g. PR-A – NDCR, M-S – NDCR, PR-S – NMI) and frame based (M-F – PR-F) measures. However, in some cases the correlations between a segment and a frame based measure are higher in this case than when adding segments.

The analysis of the precision/recall values on the simulated results (with all transformations applied) shows that there are relatively high correlations among these measures. Of the other measures NMI correlates better with recall, while M-F and NDCR have higher correlation with precision. These observations can be explained as follows. Due to the fact that NMI is normalized by the entropy of the result and truth clusters, NMI is sensitive to reducing the size of one of the cluster sets. If the size of both cluster sets remains the same, but only the size of the overlapping set of segments is changed, NMI correlates well with both precision and recall. This is the case when the segments of the simulated results are shifted, where the correlation coefficients between precision and NMI are in the range 0.53 to 0.65, and in the range 0.53 to 0.77 for recall. NMI does not distinguish between result and truth set, but just takes their overlap into account. The correlation to recall is better, as the value of NMI is mainly determined by the size of the overlapping data set and not so much biased by the size of the result set. The correlations of NMI to precision and recall are similar on the algorithm results, but the differences are not so salient there. The low (and partly negative) cor-

**Table 5: Correlation of precision (upper table) and recall (lower table) of measures on the simulated results generated by applying all transformations (Pearson’s product moment correlation coefficients). Coefficients with a confidence interval  $p > 0.005$  are written italic.**

	PR-S	PR-F	NMI	M-S	M-F	NDCR
<i>Precision</i>						
PR-A	0.633	0.622	0.150	0.746	0.512	-0.692
PR-S		0.583	0.230	0.584	0.426	-0.491
PR-F			0.096	0.511	0.775	-0.570
<i>Recall</i>						
PR-A	0.533	0.719	0.394	0.772	0.128	<i>-0.020</i>
PR-S		0.603	0.593	0.269	-0.080	0.047
PR-F			0.504	0.460	0.043	<i>0.042</i>

relation of M-F to recall is due to the fact that the measure becomes negative, if the number of missed frames is larger than the number of frames in the ground truth. This is the case when the number of frames in the result is much higher than that of the ground truth, which typically happens in cases of high recall and low precision. On the algorithm results this effect does not occur, as the algorithm results have typically much higher precision than recall. In NDCR,  $P_{miss}$  is related to recall, and the false alarm rate is related to precision. The false alarm rate is numerically much higher on the simulated results and thus dominates the overall measure. When by changing the expected false alarm rate or costs  $\beta$  is set to e.g. 0.001, the correlations with precision range from  $-0.37$  to  $-0.55$  and those with recall from  $-0.40$  to  $-0.88$ . On the results of the algorithm, the precision is typically much higher than the recall, thus NDCR has also a better correlation with recall for  $\beta = 2$ .

The experiment analyzing the correlation with human judgment of repeated content shows that it is low for all of the measures. PR-F performs slightly better than the others. As for 4 out of 6 measures rank correlation is (slightly) higher than product moment correlation, it seems that it is easier to follow the human perception in relative ranking than in absolute values. Some of the videos contain clips that are technically not near duplicates, but do not provide much new information to the viewer. It seems that humans judge this more holistic impression of repeated content, i.e. the amount of “redundancy” or “boringness”. It also has to be noted that due to the fact that the videos are summaries, they contain a number of short segments, the results might be different on longer videos with a slower editing pace.

## 4. CONCLUSION

We have surveyed the different measures proposed for evaluating detection of near duplicate videos and analyzed their correlation on real algorithm results as well as on simulated results. The measures make different assumptions about the inputs, e.g. same segmentation of result set and ground truth, and measure on different granularity (segment, frame). The results show that depending on the type of differences (segments added, dropped, shifted) between ground truth and results the correlation between the measures can in some cases be quite low, so that results obtained from different measures cannot be compared as it is sometimes found in literature. Shifting segment boundaries has the strongest impact on the correlation of the measures, and does not only affect frame based measures, but also shot

based ones that assume alignment between result and truth segments.

In general we find no grouping into frame and segment based measures. However, the different variants of precision/recall type measures have higher correlations among them than others. Also the shot based Muscle VCD measure is well correlated to them. The TRECVID NDCR measure does also correlate if the costs and the expected false alarm rate are chosen for the specific application. Some measures such as NMI and the frame based Muscle VCD measure have issues when the number of segments in the result set and the ground truth change strongly. The segment based precision/recall measure is not robust in cases with no or a very large number of near duplicates.

For obtaining comparable evaluation results, choosing one of the measures from the well correlated group seems to be advantageous. The TRECVID NDCR measure needs to be parameterized for a certain application scenario, which might hinder comparability across applications. The aligned cluster and frame based precision/recall measures both do not require the same segmentation on ground truth and result set and support clustering. These are useful properties for many practical problems. Depending on the specific problem, frame precise results may be available or even required. In this case the frame based precision/recall measure seems to be most appropriate, otherwise the aligned cluster precision/recall measure can be used.

The correlation of the measures with human perception is generally weak, but rank correlation is slightly higher. However, the frame based precision/recall measure performs best. Human perception seems to take a broader range of factors into account than covered by near duplicates. This issue needs to be explored further on a more diverse content set than the summary videos, with the goal to develop measures that correlate better with human perception. Such measures would significantly support the development of near duplication detection algorithms that determine redundancy on a higher semantic level than today’s algorithms.

## Acknowledgments

The author would like to thank Georg Thallinger and Werner Haas for their support. The research described here has been partially supported by the European Commission under the contract FP7-215475, “2020 3D Media” (<http://www.20203dmedia.eu/>).

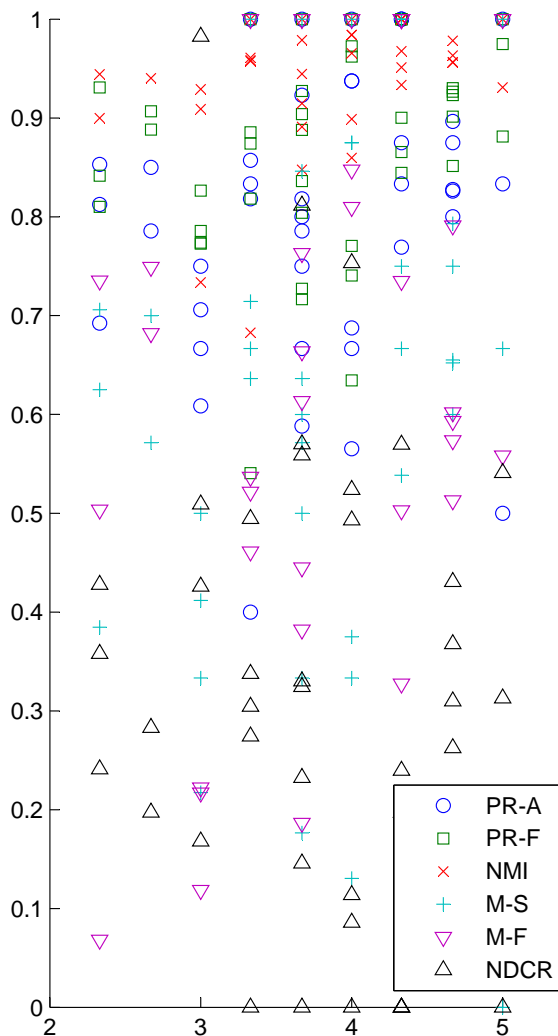


Figure 1: Scatter plot of (F1) measures (y-axis) against human evaluator judgments (x-axis) of the amount of repeated content in summary videos.

## 5. REFERENCES

- [1] W. Bailer, F. Lee, and G. Thallinger. Skimming rushes video using retake detection. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 60–64, Sept. 2007.
- [2] W. Bailer, F. Lee, and G. Thallinger. Detecting and clustering multiple takes of one scene. In *Proceedings of 14th Multimedia Modeling Conference*, pages 80–89, 2008.
- [3] W. Bailer, F. Lee, and G. Thallinger. A distance measure for repeated takes of one scene. *The Visual Computer*, 25(1):53–68, Jan. 2009.
- [4] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68, New York, NY, USA, 2004. ACM.
- [5] M. Bertini, A. D. Bimbo, and W. Nunziati. Video clip matching using MPEG-7 descriptors and edit distance. In *CIVR*, pages 133–142, 2006.
- [6] E. Dumont and B. Mérialdo. Rushes video parsing using video sequence alignment. In *CBMI 2009, 7th International Workshop on Content-Based Multimedia Indexing, June 3-5, 2009, Chania, Crete Island, Greece, Jun. 2009*.
- [7] P. Duygulu, J.-Y. Pan, and D. A. Forsyth. Towards auto-documentary: tracking the evolution of news stories. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 820–827, New York, NY, USA, 2004. ACM.
- [8] J. Law-To, A. Joly, and N. Boujemaa. Muscle-VCD-2007: a live benchmark for video copy detection, 2007. <http://www-rocq.inria.fr/imedia/civr-bench/>.
- [9] C.-W. Ngo, W.-L. Zhao, and Y.-G. Jiang. Fast tracking of near-duplicate keyframes in broadcast domain with transitivity propagation. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 845–854, New York, NY, USA, 2006. ACM.
- [10] NHK Science & Technical Research Laboratories. Test modules for TRECVID activity. Use case scenario. Ver.1.2.0E, Apr. 2008.
- [11] P. Over, A. F. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 1–15, New York, NY, September 2007. ACM Press.
- [12] S. Poullot, M. Crucianu, and O. Buisson. Scalable mining of large video databases using copy detection. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 61–70, New York, NY, USA, 2008. ACM.
- [13] E. Rossi, S. Benini, R. Leonardi, B. Mansencal, and J. Benois-Pineau. Clustering of scene repeats for essential rushes preview. *Image Analysis for Multimedia Interactive Services, International Workshop on*, 0:234–237, 2009.
- [14] S. Satoh, M. Takimoto, and J. Adachi. Scene duplicate detection from videos based on trajectories of feature points. In *MIR '07: Proceedings of the international workshop on multimedia information retrieval*, pages 237–244, New York, NY, USA, 2007. ACM.
- [15] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.
- [16] M.-C. Yeh and K.-T. Cheng. Video copy detection by fast sequence matching. In *ACM International Conference on Image and Video Retrieval*, July 2009.