

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

J.244

(04/2008)

SERIES J: CABLE NETWORKS AND TRANSMISSION
OF TELEVISION, SOUND PROGRAMME AND OTHER
MULTIMEDIA SIGNALS

Measurement of the quality of service

**Full reference and reduced reference calibration
methods for video transmission systems with
constant misalignment of spatial and temporal
domains with constant gain and offset**

Recommendation ITU-T J.244

Recommendation ITU-T J.244

Full reference and reduced reference calibration methods for video transmission systems with constant misalignment of spatial and temporal domains with constant gain and offset

Summary

Recommendation ITU-T J.244 specifies full reference (FR) and reduced reference (RR) calibration methods which report alignment shifts in both space and time of processed (e.g., compressed, transmitted, and decompressed) video sequences with respect to corresponding source video sequences. The methods also compute gain and offset of processed video sequences. The methods require that spatial and temporal shifts must be constant. Furthermore, the gain and offset must be constant. The methods in this Recommendation are applicable to:

- reporting shifts that can be used to register source and processed video sequences in space and time;
- reporting gains, and offsets of video levels that can be used to calibrate processed video sequences with respect to the source.

Intended applications for these methods include, but are not limited to, assessing the operational readiness of video transmission systems that contain compression/transmission/decompression components, obtaining calibration estimates for correcting processed video sequences before measuring subjective or objective picture quality, and verifying that video transmission systems meet specified calibration requirements.

Source

Recommendation ITU-T J.244 was approved on 6 April 2008 by ITU-T Study Group 9 (2005-2008) under Recommendation ITU-T A.8 procedure.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure e.g. interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2009

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

CONTENTS

Page

1	Scope	1
1.1	Applications.....	1
1.2	Limitations.....	2
2	References.....	2
2.1	Normative references.....	2
2.2	Informative references.....	2
3	Abbreviations.....	2
4	Summary of validation results	3
5	Description of the calibration methods.....	5
Annex A	– Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce – Reduced Reference Video Calibration Algorithms	6
A.1	Introduction	6
A.2	Temporal registration algorithm.....	6
A.3	Spatial registration algorithm	14
A.4	Valid region algorithm	26
A.5	Luminance gain and offset algorithm.....	30
A.6	Combining algorithms and applying corrections	33
A.7	References	34
Annex B	– Yonsei University – A calibration method for constant misalignment of spatial and temporal domains with constant gain and offset.....	35
B.1	Introduction	35
B.2	Spatial and temporal registration.....	35
B.3	Gain and offset estimation.....	36
B.4	Refinement	38
Appendix I	– Example Validation Testplan for Calibration Methods	39
I.1	Introduction	39
I.2	Example Calibration test plan (no variable temporal misalignment)	39
I.3	Validation process	40
I.4	Generation of Processed Video Sequences (PVSs)	40
I.5	Evaluation metrics	41
I.6	Additional information	41
I.7	Future test plans (variable temporal misalignment for multimedia applications)	41
Appendix II	– Test Plan 1 (Rec. 601 only – Jun 2, 2007).....	42
II.1	Calibration test plan 1 (no variable temporal misalignment)	42
II.2	Validation process	42
II.3	Generation of Processed Video Sequences (PVSs)	42

	Page
II.4 Evaluation metrics	43
Appendix III – Evaluation Metrics	44
III.1 Introduction	44
III.2 Experimental results of 1190 PVSs for each video format.....	44
III.3 Experimental results after removing erroneous video sequences.....	47
III.4 Experimental results per HRC after removing erroneous video sequences	50

ITU-T J.244 (04/2008)

Introduction

Calibration techniques described in this Recommendation may be used for various purposes. Some of these include the verification of system setup, maintenance of delivery systems, and pre-processing for some objective video quality assessment techniques.

There has been a great interest in objective models which automatically measure video quality. For some objective quality assessment methods, the source and processed (e.g., compressed, transmitted, and decompressed) video sequences need to be registered. It is noted that some application or service scenarios may introduce spatio-temporal misalignment between processed and source video sequences. In the case where such spatio-temporal alignment problems occur, the prediction accuracy of some objective quality assessment models can be reduced. Typically, for full reference and some reduced reference objective quality assessment models, it is important that the processed and reference pictures are properly aligned. For some objective methods, alignment can be an integral part of the model. In such models there is no need for an external alignment method. For objective models that do not include inherent alignment, there may be the need for a separate picture alignment procedure that can be performed before parameter values are derived.

Intended applications for these methods include, but are not limited to, assessing the operational readiness of video transmission systems that contain compression/transmission/decompression components, obtaining calibration estimates for correcting processed video sequences before measuring subjective or objective picture quality, and verifying that video transmission systems meet specified calibration requirements.

Copyright © 2008 ITU-T. All rights reserved.

Copyright International Telecommunication Union/ITU Telecommunication Sector
Provided by IHS Markit under license with ITU-T
No reproduction or networking permitted without license from IHS

Recommendation ITU-T J.244

Full reference and reduced reference calibration methods for video transmission systems with constant misalignment of spatial and temporal domains with constant gain and offset

1 Scope

This Recommendation specifies calibration methods which report alignment shifts in both space and time of processed (e.g., compressed, transmitted, and decompressed) video sequences with respect to corresponding source video sequences. The methods also compute gain and offset of processed video sequences. The methods require that spatial and temporal shifts must be constant. Furthermore, the gain and offset must be constant. The methods in this Recommendation are applicable to:

- reporting shifts that can be used to register source and processed video sequences in space and time;
- reporting gains, and offsets of video levels that can be used to calibrate processed video sequences with respect to the source.

These methods can be used to report calibration differences between the source and processed video sequences. These values may be used to repair processed video sequences.

Table 1 summarizes the required bandwidth and Table 2 shows the ranges of spatial and temporal shifts for which the methods in this Recommendation were tested.

Table 1 – Allowed bandwidth

	Available bandwidth for registration					
ITU-R 601(interlaced)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s
VGA (progressive)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s
CIF (progressive)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s
QCIF (progressive)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s
HD (interlaced)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s
HD (progressive)	Unlimited	56 kbit/s	10 kbit/s	5 kbit/s	1 kbit/s	0.5 kbit/s

Table 2 – Expected conditions

Source video length	8 seconds
Maximum spatial shift (vertical and horizontal)	±20 pixels
Maximum temporal shift	±1 second
Gain (Y-component)	±10%
Offset (Y-component)	±10 grey levels

1.1 Applications

Intended applications for these methods include, but are not limited to, assessing the operational readiness of video transmission systems that contain compression/transmission/decompression components, obtaining calibration estimates for correcting processed video sequences before measuring subjective or objective picture quality, and verifying that video transmission systems meet specified calibration requirements.

1.2 Limitations

The methods presented in this Recommendation require access to both source and processed video sequences. The spatial and temporal shifts must remain constant for the entire processed video sequences. Furthermore, the gain and offset must be fixed.

NOTE – The structure and content of this Recommendation have been organized for ease of use by those familiar with the original source material; as such, the usual style of ITU-T recommendations has not been applied.

2 References

2.1 Normative references

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T J.144] Recommendation ITU-T J.144 (2004), *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*.

[ITU-R BT.1683] Recommendation ITU-R BT.1683 (2004), *Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference*.

2.2 Informative references

None.

3 Abbreviations

This Recommendation uses the following abbreviations:

CODEC	CODer-DECoder
FR	Full Reference
GOP	Group Of Pictures
HRC	Hypothetical Reference Circuit
IRT	Institut Rundfunk Technische (Germany)
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
NR	No (or Zero) Reference
NTSC	National Television Standard Committee (60 Hz TV)
PAL	Phase Alternating Line standard (50 Hz TV)
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit

SSCQE Single Stimulus Continuous Quality Evaluation

VQEG Video Quality Experts Group

4 Summary of validation results

In accordance with the validation testplan for calibration methods in Appendices I and II, a total of 1190 test video sequences (525 video format) are generated from the VQEG Phase I data for various test conditions (different source video sequences and HRCs). The evaluation metrics of the testplan are provided in Appendix III. However, it was found that some of the original VQEG Phase I data were not correctly calibrated and some video sequences include still scene. Consequently, these incorrectly-calibrated and still-scene video sequences will inevitably introduce additional errors which have nothing to do with the calibration methods. When these problematic video sequences are removed, there are 973 PVSs for the 525 video format and 812 PVS for the 625 video format. Furthermore, since the test video sequences are stored in the integer format, they inherently include round-up errors which no model can predict. With those round-up errors, it would be meaningless to present the percentiles of correct results with no error in any aspect.

Tables 3-6 show the mean values of the five errors and error statistics for the 525 video format after removing the problematic video sequences. Tables 7-8 show the mean values of the five errors and error statistics for the 625 video format after removing the problematic video sequences. Appendix III presents additional error statistics.

Table 3 – Mean values of the five errors (525 format, 973 PVSs)

	Mean (Ex)	Mean (Ey)	Mean (Et)	Mean (Eg)	Mean (Eo)
NTIA 25k	0	0	−0.015	0.005	1.111
Yonsei unlimited	0	−0.00084	0.006723	0.005649	0.621204
Yonsei 56k	0	0.007563	0	0.005799	0.62936
Yonsei 10k	−0.00168	−0.00168	0.014286	0.006	0.624177
Yonsei 5k	0.00084	0.007563	0.002521	0.006184	0.590344
Yonsei 1k	0.002521	−0.00252	0.021008	0.014237	−0.56031
Yonsei 0.5k	−0.00924	0.010084	0.058824	0.01666	−0.68251

Table 4 – Error statistics of mean values (525 format, 973 PVSs)

Percentile	Ex+Ey≤0	Ex+Ey≤1	Ex+Ey+Et=0	Ex+Ey+Et≤1	Ex+Ey+Et≤2
NTIA 25k	100	100	98.739	99.748	100
Yonsei unlimited	99.91597	100	99.41176	99.83193	100
Yonsei 56k	97.22689	99.4958	96.55462	98.40336	99.91597
Yonsei 10k	99.07563	99.91597	98.23529	99.2437	99.91597
Yonsei 5k	96.38655	99.2437	94.70588	97.31092	99.4958
Yonsei 1k	95.79832	99.2437	93.61345	96.80672	99.32773
Yonsei 0.5k	95.04202	98.90756	91.84874	95.12605	97.73109

Table 5 – Supplementary error statistics of mean values (525 format, 973 PVSs)

Percentile	Et≤0	Et≤1	Et≤2	Eg≤0.01	Eg≤0.02	Eo≤2.5	Eo≤5.0
NTIA 25k	98.739	99.748	100	82.353	89.916	82.353	94.118
Yonsei unlimited	99.41176	99.91597	100	83.10924	95.71429	91.59664	98.48739
Yonsei 56k	98.23529	99.91597	100	82.26891	94.28571	90.84034	97.89916
Yonsei 10k	98.48739	99.91597	100	81.59664	93.52941	90.2521	97.89916
Yonsei 5k	96.47059	99.57983	99.7479	80.58824	92.43697	90.42017	97.47899
Yonsei 1k	95.88235	99.15966	99.66387	51.42857	75.12605	82.10084	94.78992
Yonsei 0.5k	93.94958	98.23529	98.7395	47.14286	69.15966	79.41176	93.44538

Table 6 – Mean values of the five errors (625 format, 812 PVSs)

	Mean (Ex)	Mean (Ey)	Mean (Et)	Mean (Eg)	Mean (Eo)
NTIA 25k	−0.042016807	−0.058823529	0.304201681	0.003310455	0.270601479
Yonsei unlimited	0.061345	0.018487	0.003361	0.005358	0.350661
Yonsei 56k	0.056303	0.022689	0.012605	0.005665	0.483693
Yonsei 10k	0.052941	0.026891	−0.12101	0.010251	0.901221
Yonsei 5k	0.045378	0.017647	−0.1	0.010353	0.864778
Yonsei 1k	0.086555	0.022689	−0.23109	0.022666	−1.38218
Yonsei 0.5k	0.291597	0.021008	−0.30756	0.036449	−2.527

Table 7 – Error statistics of mean values (625 format, 812 PVSs)

Percentile	Ex+Ey≤0	Ex+Ey≤1	Ex+Ey+Et=0	Ex+Ey+Et≤1	Ex+Ey+Et≤2
NTIA 25k	92.43697	98.31933	80	84.03361	90.08403
Yonsei unlimited	94.45378	99.07563	84.36975	86.89076	90.33613
Yonsei 56k	94.03361	98.82353	83.94958	86.21849	89.74790
Yonsei 10k	93.52941	98.57143	84.11765	85.96639	90.50420
Yonsei 5k	93.94958	98.48739	84.20168	86.47059	90.08403
Yonsei 1k	93.44538	97.64706	83.78151	86.05042	88.90756
Yonsei 0.5k	90.42017	95.37815	80.58824	83.94958	86.55462

Table 8 – Supplementary error statistics of mean values (625 format, 812 PVSs)

Percentile	Et≤0	Et≤1	Et≤2	Eg≤0.01	Eg≤0.02	Eo≤2.5	Eo≤5.0
NTIA 25k	80	88.82352	91.428571	88.235294	97.47899	96.63865	100
Yonsei unlimited	85.96639	90.84034	91.26050	86.63866	92.94118	93.02521	96.38655
Yonsei 56k	85.46218	90.58824	90.92437	83.44538	92.18487	92.60504	95.71429
Yonsei 10k	85.71429	91.0084	91.76471	77.05882	86.21849	85.88235	90.42017
Yonsei 5k	86.05042	90.7563	91.59664	77.22689	86.13445	86.38655	91.00840
Yonsei 1k	85.71429	90.84034	91.2605	39.41176	58.9916	64.87395	84.87395
Yonsei 0.5k	83.52941	89.83193	90.7563	32.01681	46.63866	56.55462	76.30252

5 Description of the calibration methods

The calibration methods are described in Annexes A and B.

Annex A

Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce

Reduced Reference Video Calibration Algorithms

(This annex forms an integral part of this Recommendation)

This annex describes four Reduced Reference (RR) video calibration algorithms of low computational complexity. RR methods are useful for performing end-to-end in-service video quality measurements since these methods utilize a low bandwidth network connection between the original (source) and processed (destination) ends. The first RR video calibration algorithm computes temporal registration of the processed video stream with respect to the original video stream (i.e., video delay estimation). The second algorithm jointly calculates spatial scaling and spatial shift. The third algorithm calculates luminance gain level offset of the processed video stream with respect to the original video stream. The fourth algorithm estimates the valid video region of the original or processed video stream (i.e., the portion of the video image that contains actual picture content). All the algorithms utilize only the luminance (Y) image plane of the video signal.

Keywords: calibration; delay; gain; offset; spatial scaling; spatial shift; temporal shift; video.

A.1 Introduction

This annex describes a series of low computational complexity Reduced Reference (RR) video calibration algorithms. RR measurements are useful for in-service quality monitoring applications since they require only a small amount of reference information to make a performance measurement [1]. This RR information can thus be easily communicated between the original (i.e., source) and processed (i.e., destination) ends of a video system using commonly available network connections (e.g., Internet, Public Switched Telephone Network). In an RR measurement system, video calibration of the processed video stream (i.e., estimating gain, offset, temporal delay, spatial shift, spatial scaling, and valid region) is normally a prerequisite for estimating the quality of the processed video stream. While the algorithms described in this Recommendation can produce calibration estimates using a single video clip, these algorithms are most effective when applied in the order specified to a series of video clips associated with a single video system. Combining calibration results from multiple video clip samples provides a means to obtain more robust estimates.

The first algorithm estimates the overall temporal alignment (i.e., delay) of the processed video stream with respect to the original video stream. This temporal delay estimation algorithm utilizes three features that characterize the motion and luminance of the video scene. The second algorithm jointly estimates spatial scaling and spatial shift. This second algorithm uses randomly selected pixels and image profiles (horizontal and vertical) that are extracted from the video scene. The third algorithm reuses these pixels and profiles to estimate luminance gain and level offset. The fourth algorithm examines the edges of images and estimates the portion of the video image (original or processed) that contains actual valid picture content. All algorithms utilize the luminance (Y) image plane of the video signal. Luminance values are presumed to range from 1 to 254 (e.g., black = 16, white = 235).

A.2 Temporal registration algorithm

The temporal registration algorithm presented in this clause is an RR method that requires a data transmission bit-rate of 1 to 6 kbit/s (depending upon video frame rate and feature quantization

accuracy¹). This low bit-rate requirement, together with the computational simplicity of extracting the RR features, makes this method ideally suited for real-time in-service monitoring of end-to-end video delay. Three low bandwidth features track changes in scene motion and image intensity. Video delay is estimated by cross-correlating, or aligning, the information in the processed feature streams with the corresponding information in the original feature streams. This procedure produces one video delay estimate per video clip, which should have a minimal duration of 5 seconds².

A.2.1 Default valid video region

NTSC (525-line) and PAL (625-line) video sampled according to ITU-R BT.601 (henceforth abbreviated Rec. 601) may have a border of pixels and lines that do not contain valid picture. The original video from the camera may only fill a portion of the Rec. 601 frame. Some digital video compression schemes further reduce the area of the picture in order to save transmission bits. To prevent non-picture areas from degrading the performance of the temporal registration algorithm, they must be excluded.

Clause A.4 presents an automated method for determining valid region. However, since valid region is always referenced to the original video, we first need to know the spatial shift and scaling that is present in the processed video. Since spatial shift and scaling information is only available after the spatial registration algorithm is completed, and since temporal registration must be known before spatial registration, we must use some reasonable default values for valid region. Table A.1 gives the reasonable default values that we use for the border of invalid pixels around the edge of common image sizes. Pixels in this invalid border region will be discarded by the temporal registration algorithm. Images in common intermediate format (CIF), source input format (SIF), and quarter resolution versions of these (QCIF and QSIF) typically do not have an invalid border, so no pixels are discarded.

Table A.1 – Default invalid border for common video sizes

Video Type	Rows	Columns	Invalid Top	Invalid Left	Invalid Bottom	Invalid Right
NTSC (525-line)	486	720	18	22	18	22
PAL (625-line)	576	720	14	22	14	22

A.2.2 Description of Features

The temporal alignment algorithm uses three low-bandwidth features. The first temporal information (TI) feature, designated TI2, uses temporal differences between adjacent luminance images. For interlace video, fields spaced 2 fields apart in time (i.e., 1 frame apart) are differenced, and for progressive video, frames spaced one frame apart in time are differenced. This TI feature quantifies the amount of scene motion by summing the squared pixel differences between two sampled images. The second feature, designated TI10, is based on temporal differences spaced 10 interlaced fields or 5 progressive frames apart. If the frame rate in units of frames per second (FPS) decreases and the digital video system fills in the non-transmitted frames with repetitions of prior frames, processed scene motion is perceived as unnatural or discontinuous. The second feature smoothes over these times of no motion in the processed video (i.e., repeated frames or fields). The

¹ 6 kbit/s presumes interlaced video operating at 60 fields per second and single precision to code the features.

² The method presented here estimates the average video delay of the entire processed sequence, as opposed to methods that estimate the video delays of individual frames within the processed sequence. When individual frames within the processed sequence have different video delays, they will tend to vary about the average video delay found by this algorithm.

third feature, designated Y_{mean} , is based on the mean of the Y images. This feature quantifies the intensity, or brightness, of the video scene by measuring the average luminance level.

Each of the three features characterizes the video differently. Thus, each works best for different amounts of frame repeating, varying video delay, and scene intensity fluctuations; and one feature may produce good temporal alignments where another one fails. The magnitude and shape of the correlation function are used as indicators for the reliability of that feature's temporal registration. Unreliable features are discarded, and the remaining results are considered jointly to estimate the best average video delay for the clip.

In the following feature definitions, the luminance image is noted as Y . For interlaced video, Y is a field and for progressive video, Y is a frame. The time when this field or frame occurs is denoted as t . Pixels of Y are further subscripted by row and column, i and j , respectively, so that an individual pixel is denoted as $Y(i,j,t)$.

A.2.2.1 TI2 Feature: Two Field Difference Temporal Information

For interlaced video, to compute TI2 at time t , consider field $Y(t)$ and the previous field of the same type $Y(t-2)$, and compute for each pixel

$$TI2(i, j, t) = Y(i, j, t) - Y(i, j, t-2) \quad (1)$$

For progressive video, consider frame $Y(t)$ and the previous frame $Y(t-1)$, and compute for each pixel

$$TI2(i, j, t) = Y(i, j, t) - Y(i, j, t-1) \quad (2)$$

Then, using the results from (1) or (2), compute

$$TI2(t) = rms_{\text{space}}[TI2(i, j, t)] = \sqrt{\frac{1}{R} \sum_i \sum_j (TI2(i, j, t))^2} \quad (3)$$

where rms_{space} is the root mean square function over space defined by the above equation, i and j are within the valid region defined in clause A.2.1, and R is the total number of pixels in the valid region (i.e., the number of pixels in the double summation). The TI2 calculation for interlaced fields is calculated as shown in Figure A.1.

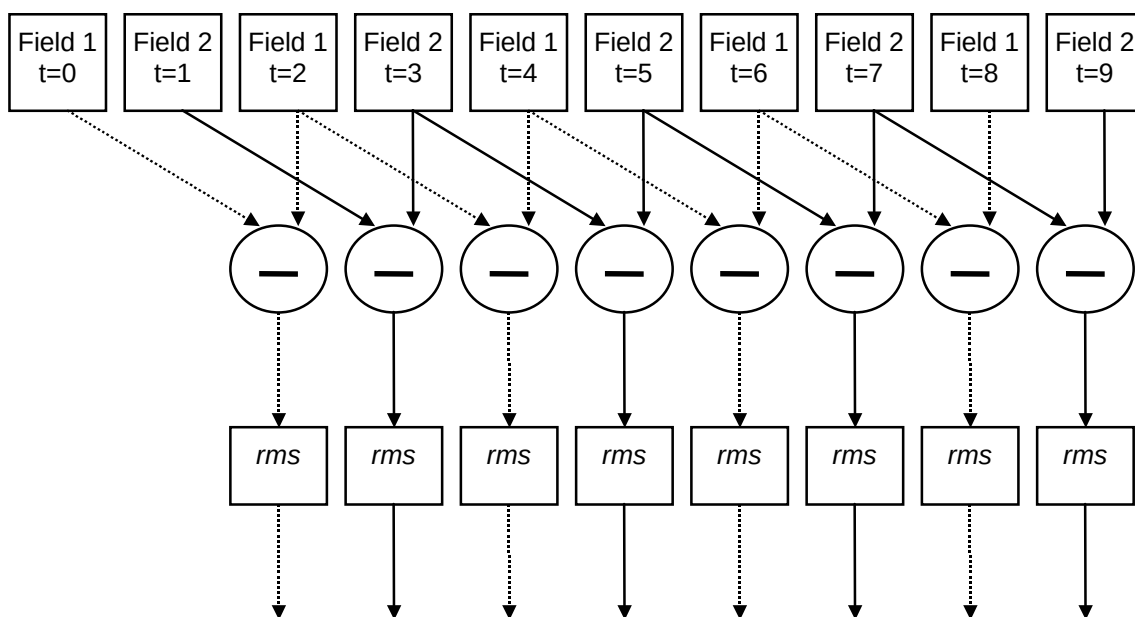


Figure A.1 – Diagram depicting the method of calculating TI2(t)

A.2.2.2 TI10 Feature: Ten field difference temporal information

The TI10 feature is based on a temporal difference spaced ten interlaced fields apart, or five progressive frames apart. This feature smoothes the temporal information using a wider filter than TI2 and eliminates the effect of frame repeats in the TI waveform for systems that have four or fewer consecutive frame repeats.

To compute TI10 at time t on interlaced sequences, consider field $Y(t)$ and the field of the same type five frames ago $Y(t-10)$, and compute for each pixel

$$TI10(i, j, t) = Y(i, j, t) - Y(i, j, t - 10) \quad (4)$$

For progressive sequences, consider frame $Y(t)$ and the frame five frames ago $Y(t-5)$, and compute for each pixel

$$TI10(i, j, t) = Y(i, j, t) - Y(i, j, t - 5) \quad (5)$$

Then, using the results from (4) or (5), compute

$$TI10(t) = rms_{space}[TI10(i, j, t)] \quad (6)$$

where i and j are within the valid region defined in clause A.2.1.

A.2.2.3 Ymean feature: average luminance level

Ymean is calculated as the average luminance level of a field. To compute Ymean at time t on interlaced fields or frames, consider $Y(t)$ and compute

$$Ymean(t) = mean_{space}[Y(i, j, t)] \equiv \frac{1}{R} \sum_i \sum_j Y(i, j, t) \quad (7)$$

where $mean_{space}$ is the mean function over space defined by the above equation, i and j are within the valid region defined in clause A.2.1, and R is the total number of pixels in this valid region (i.e., the number of pixels in the double summation).

A.2.3 Feature sequence cross-correlation³ and validation

The following steps describe the process that is used to cross-correlate original and processed feature streams in order to estimate their best temporal registration. The algorithm includes validation steps, where the features are examined to discard potentially unreliable alignment results. The original feature stream will be referred to as a_o and the processed feature stream as a_p . For example, when using the two field difference temporal information feature, $a(t) = TI2(t)$. Let M be the length of the feature streams, a_o and a_p . The feature stream a_o , given by $\{a_o(0), a_o(1), a_o(2), \dots, a_o(M-1)\}$ will be abbreviated as $\{a_o(t)\}_{t=0}^{M-1}$.

We have found that temporal registration estimates based on the above features become unreliable when

$$std_{time}[a_o(t)] \leq threshold \quad (8)$$

or

$$std_{time}[a_p(t)] \leq threshold \quad (9)$$

³ The correlation function described in this clause is based on minimization of the standard deviation of the difference between the original and processed feature streams, not the maximization of the energy in the cross-product of the two feature streams.

where $threshold = Y_THRESHOLD = 0.25$ for the Ymean feature, $threshold = TI_THRESHOLD = 0.15$ for the TI2 and TI10 features, and std_{time} represents the standard deviation over time of the M samples in the feature stream.

When (8) or (9) are satisfied, the Ymean waveform has detected insufficient temporal changes in scene brightness to be useful for temporal registration (e.g., scene with a constant brightness level). Similarly when (8) or (9) are satisfied the TI2 and TI10 waveforms have detected insufficient temporal changes in the scene motion to be useful for temporal registration (e.g., still scene). Features that fall below the thresholds in (8) or (9) are considered "invalid" and no further calculations are performed using them. Furthermore, if all three features (TI2, TI10, and Ymean) fall below the thresholds in (8) or (9), then the video clip is considered "still" and no further temporal registration calculations are performed.

Feature sequence correlation is performed on the feature streams that pass the above test. The temporal registration uncertainty, U, will be specified in *fields* for interlaced systems and *frames* for progressive systems. U represents the maximum temporal shift (plus or minus) of the processed feature stream with respect to the original feature stream. The feature sequence correlation is performed as follows:

- 1) Given a sequence of M processed video features $\{a_p(t)\}_{t=1}^M$, we first discard the first and last U samples to form sequence $\{a_p(t)\}_{t=U}^{M-1-U}$. Normalize (divide) each element in the resulting sequence by the standard deviation of that sequence to form the normalized sequence $\{n_p(t)\}_{t=U}^{M-1-U}$.
- 2) For each alignment delay guess d (for all $-U \leq d \leq U$), we compute a corresponding original feature stream $\{a_o(t)\}_{t=U+d}^{M-1-U+d}$, and normalize (divide) each element in the sequence by the standard deviation of that sequence to form the normalized sequence $\{n_o(d,t)\}_{t=U}^{M-1-U}$. This original feature normalization is essentially the same as the processed feature normalization in step one, except computed for each delay, d .
- 3) Take the resulting normalized processed and original feature streams and compute the difference between those sequences:

$$Diff(d,t) = \{n_o(d,t) - n_p(t)\}_{t=U}^{M-1-U} \quad (10)$$

- 4) Compute the sample standard deviation over time of the difference sequence for each delay offset d , namely

$$S(d) = std_{time}(Diff(d,t)) \quad (11)$$

- 5) The minimum $S(d)$ (denoted S_{min}) and its offset d_{min} is the best alignment indicated by this feature. Figure A.2 gives an example plot of the correlation function $S(d)$. In this plot, the best alignment occurs for the delay $d = 0$ fields (i.e., $d_{min} = 0$ fields).

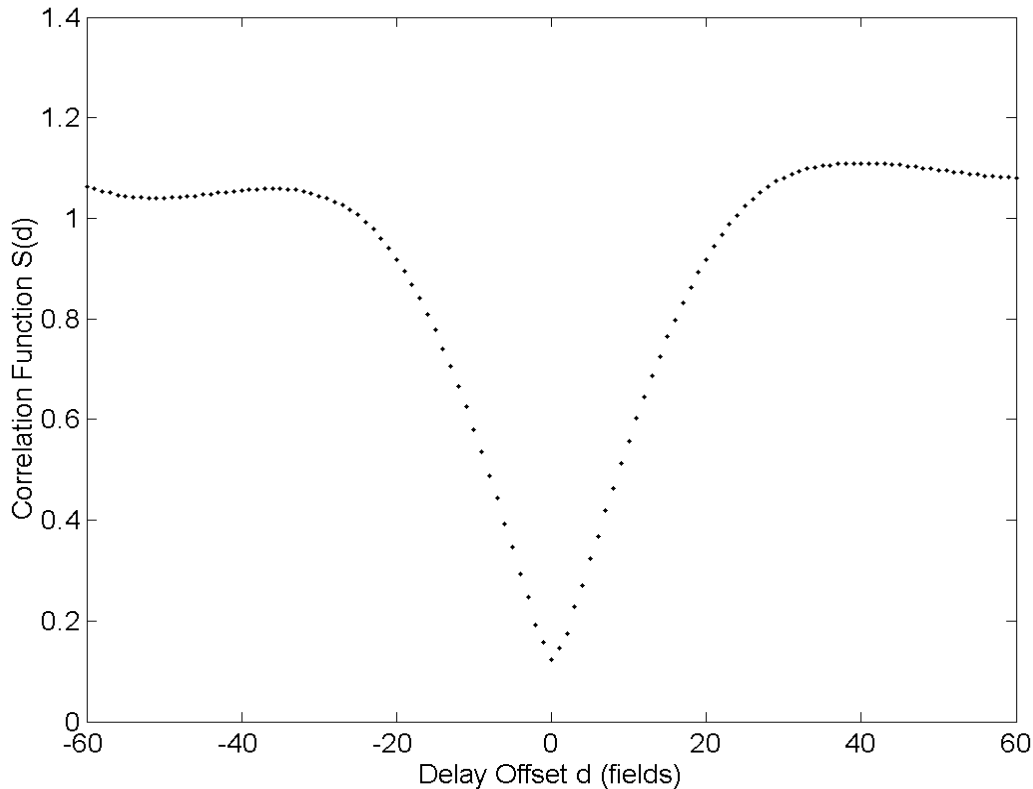


Figure A.2 – Example plot of correlation function $S(d)$

- 6) If the normalized original and processed feature streams were identical, then they would cancel at correct alignment (i.e., $S_{\min}$ would be 0.0). On the other hand, $S_{\min}$ can be at most 2.0, since the normalized original and processed waveforms each have a variance of 1.0. If the normalized original and processed waveforms are independent, their variances will add and $S_{\min}$ will be approximately equal to $\sqrt{2} \approx 1.414$. A value for $S_{\min}$ greater than $\sqrt{2}$ indicates that the two waveforms are negatively correlated. We have found that if $S_{\min} \leq \text{CORRELATION_VALID} = 0.25$, then the correlation between the processed feature stream and the original feature stream is probably reliable and these features are therefore considered valid. However, if $S_{\min} \geq \text{CORRELATION_INVALID} = 1.40$, then the correlation between the processed feature stream and the original feature stream is unreliable for the above reasons and these features are therefore considered invalid. For correlations in-between (e.g., $\text{CORRELATION_VALID} \leq S_{\min} \leq \text{CORRELATION_INVALID}$), $S_{\min}$ yields ambiguous results with respect to accuracy, so other criteria must be used (see step 7).
- 7) If ($\text{CORRELATION_VALID} \leq S_{\min} \leq \text{CORRELATION_INVALID}$), find the earliest (minimum) delay, d_1 , where $S(d_1) < S_{\min} + \text{DELTA_THRESHOLD}$, where $\text{DELTA_THRESHOLD} = 0.04$. Also find the latest delay, d_2 , where $S(d_2) < S_{\min} + \text{DELTA_THRESHOLD}$. Compute the distance between those two delays, $w = d_2 - d_1 + 1$. Notice that no restrictions are placed on values of S for delays between d_1 and d_2 . This width w discriminates between correlations with a well-defined minimum, and correlations with multiple nearly-identical minimum values (e.g., a sharp correlation function as given in Figure A.2 versus a broad correlation function). TI features with $w \leq \text{TI_WIDTH}$ are reliable, and Ymean features with $w \leq \text{Y_WIDTH}$ are reliable, where $\text{TI_WIDTH} = 3$ and $\text{Y_WIDTH} = 4$. Features that meet these criteria are considered valid whereas features that do not meet these criteria are considered invalid.

Table A.2 provides a summary of the recommended threshold values that are used by the temporal registration algorithm. These recommended threshold values were obtained by minimizing alignment errors between the current algorithm and the frame-based temporal alignment method in clause 3.4.2 of [3].

Table A.2 – Recommended values for thresholds used by temporal registration algorithm

Threshold	Recommended Value
TI_THRESHOLD	0.15
Y_THRESHOLD	0.25
CORRELATION_VALID	0.25
CORRELATION_INVALID	1.40
DELTA_THRESHOLD	0.04
TI_WIDTH	3
Y_WIDTH	4

A.2.4 Estimation of Temporal Offset from Features

The following describes how to apply the three features of clause A.2.1 and the correlation algorithm of clause A.2.3 to achieve the final estimate of the temporal alignment offset.

- 1) Compute the original and processed TI2 feature streams, TI10 feature streams, and Ymean feature streams in clause A.2.1. When operating in-service, transmit the original features to the processed video location. At most, single precision will be required (i.e., 4-byte floating point), and further bandwidth savings can be obtained through quantization (e.g., 16 bits per value).
- 2) For each original and processed feature stream pair, compute the correlation function $S(d)$ according to clause A.2.3 and record whether that feature is valid or invalid, and if invalid whether that feature is still (e.g., motionless).
- 3) If one or more features are valid, average those valid correlation functions together. For progressive video sequences, find the delay offset $S_{\min}$ that minimizes the averaged correlation function. For interlaced video sequences, this algorithm may be run with either field or frame accuracy. For frame accurate delay, restrict the search to either field one delays or field two delays; for field accurate delay, include both field one and field two delays in the search. Find the delay offset $S_{\min}$ that minimizes the averaged correlation function.

NOTE – This algorithm cannot be used to determine with 100% reliability whether field one of the processed video sequence best aligns to field one or field two of the original sequence (i.e., indicative of interlaced reframing by the video system under test – see clause 3.1.2 of [3]). We have found that such a determination is approximately 90% accurate. Thus, to detect reframing we recommend the use of some other, external algorithm such as the one that will be described in clause A.3. Field accurate delays from this temporal registration algorithm provide rough estimates, suitable when subsequent steps will improve those estimates.

- 4) If all of the features are invalid, a delay cannot be computed for this video clip. Furthermore, if all features have been marked still, then the clip contains still video and this extra information may be reported to the user. Temporal alignment or registration is not required to estimate quality for still scenes.

A.2.5 Observations and Conclusions

The in-service video delay measurement algorithm presented here uses a set of very low bandwidth features extracted from original and processed video sequences. This algorithm is suitable for measuring video delay in a fully automated, in-service, real-time monitoring system or for aligning