Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality

Marcus Barkowsky*†, Jens Bialkowski*, Roland Bitto†, André Kaup*
†OPTICOM GmbH, 91052 Erlangen, Germany, {mb,rb}@opticom.de
*Chair of Multimedia Communication and Signal Processing
University Erlangen-Nuremberg, 91058 Erlangen, Germany
{barkowsky,bialkowski,kaup}@lnt.de

***Abstract*— The estimation of the video quality is often performed using a full reference approach. One of the most important steps in a video quality measurement algorithm is to find the corresponding frames between the reference and the distorted video sequence. In this paper an algorithm with three steps is proposed. First, an extended version of the phase correlation is used to find candidate images with an arbitrary temporal offset, spatial scaling or spatial shift. Based on the assumption that the spatial scaling and spatial shift does not change during the sequence a set of probable parameters is selected. Finally, a maximum likelihood estimation is applied to select those temporal offsets which support the smoothest playback. A set of video sequences degraded with several distortions which are typical for multimedia scenarios are used to compare the performance to other algorithms.**

I. Introduction

The Multimedia Group of the Video Quality Experts Group (VQEG-MM) is currently evaluating the performance of different algorithms for the objective measurement of the video quality. Typical distortions have been specified in the Multimedia Group Test Plan [1]. These distortions are seperated in spatial distortions and temporal distortions. While spatial distortions are introduced by video coding or post-processing, the temporal distortions are mainly due to deficiencies in the transmission. Often, a limited bandwidth requires a reduction of the frame rate. Additionally sometimes the playback suddenly pauses and resumes after some time because of network transmission errors. There are only a few subjective tests published which include such temporal distortions [2], [3].

The subjects in a video test perform two tasks: They implicitely and intuitively map the distorted sequence to an imagined reference signal and they base their score on the temporal distortion that they experience. The first task is discussed further in this paper.

An algorithm for the estimation of the video quality which uses the reference video signal as well as the distorted image sequence is able to model the behaviour of the viewer. For an algorithm it is even more important to correctly estimate the correspondence than for the viewer. During the quality estimation some spatial measure is calculated based on the difference between the reference and the distorted signal. If the two signals are misaligned, then a much larger spatial error results. For example, the PSNR value for the Y-component of the well-known sequence "Mobile and Calendar" drops from 54.15dB to 28.44dB if the only distortion is a temporal offset of one frame for the complete sequence.

The easiest way to estimate the correspondence is to find a matching reference frame for each distorted frame. This may be implemented independently frame by frame using different matching strategies. The authors already compared several methods in [4]. The best results were achieved using the phase correlation (PC) with additional sum of absolute difference (SAD) calculation. In this paper an enhanced version of this PC-SAD algorithm which also deals with scaling of the video sequence is proposed in Section II.

Estimating the parameters on a frame by frame basis does not lead to optimal results. Some parameters are not supposed to change during playback, especially the spatial offset and the scaling usually remain constant. This imposes additional constraints on the final match for the corresponding frame. A possible solution is to estimate several matches for each individual frame and to select the best match afterwards. In Section III the details are described.

The next step improves the result by further limiting the correspondence estimation to a smooth temporal curve. A maximum likelihood method has been developed which is similar to the Viterbi Algorithm [5]. The details are given in Section IV. Different combinations of the algorithms have been implemented and their results are shown in Section V. Finally, the conclusions are drawn in Section VI.

II. Enhanced Phase Correlation

The following method is based on an algorithm published by Reddy and Chatterji in [6]. Their excellent paper describes the use of the Fast Fourier Transform (FFT) for the registration of two images independent of a zoom factor and a rotation. Alternatively, two independent scaling values for x and y direction may be calculated. In our scenario the independent scaling is more important than a rotation because an incorrect rescaling from a picture aspect ratio of 4:3 to 16:9 occurs quite often.

In Figure 1 an overview of the complete algorithm is shown. All processing steps are only applied to the luminance component. In the first step, the images are transformed using a Discrete Fourier Transform (DFT) for each image separately.

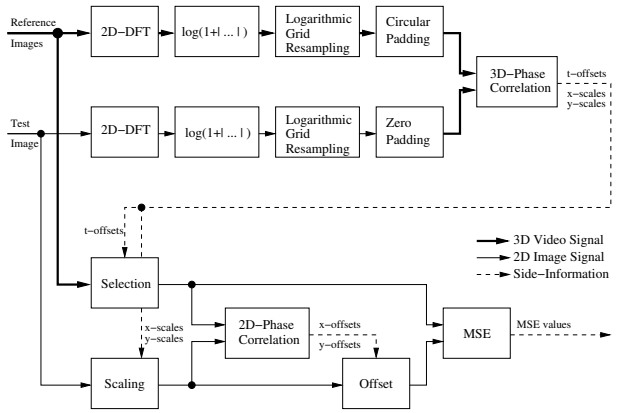

Fig. 1. Structure of parameter selection using enhanced phase correlation algorithm

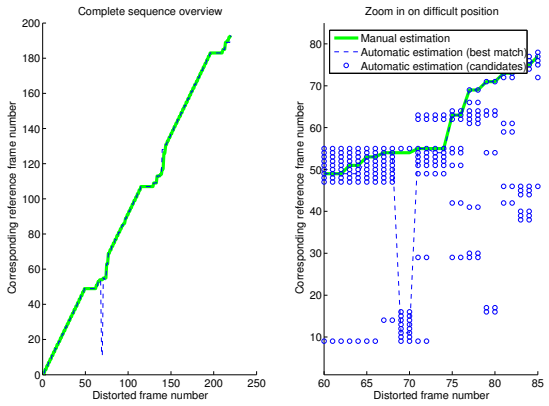

Fig. 2. Typical situation of a temporal mismatch. Left: Overview, Right: Zoom with other possible candidates shown

According to the translation property of the fourier transform, the x- and y-offset may be retrieved from the comparison of the phase of the two spectra. This implies, that the absolute value is mostly translation invariant.

For further computation only the absolute value is used. In order to emphasize the high frequency components which are usually of low energy, a logarithm is applied to the spectral values. The seperation of the image scaling is done by using the Fourier scale property and a transformation into a logarithmic coordinate system. The logarithmic coordinate system changes stretching of the spectrum into a movement which can be determined by a phase correlation step.

In our proposed algorithm this step utilizes a 3D-DFT which leads directly to the desired matching frame number as well as the scaling coefficients in x- and y-direction. There is an implementation issue however which makes a special padding step necessary. The spectral repetition of the DFT leads to a step behaviour after the last frame or before the first frame because those two frames are usually very different. The zero padding of the distorted frame leads to a similar behaviour and thus a match is often found in the 3D phase correlation. In order to avoid this, the reference sequence is circularly padded leading to the image sequence $t = 1, 2, \ldots, T-1, T, T-1, \ldots, 3, 2$. Likewise the distorted sequence is zero padded to a total length of $2T - 2$ frames.

The 3D phase correlation results in candidates for the best matching pictures and scales. The correlation value highly depends on the number of reference frames and on the characteristics of the input sequence. This is due to the energy normalization in the frequency domain which is also present in the resulting signal due to Parseval's theorem. Therefore the correlation value itself is not suited as a quantitative measure of the quality of the match found. Instead we are using the Mean Squared Error (MSE). However, some intermediate steps are necessary.

In our approach we use the position of the global maximum which is the highest correlation peak. Additionally, some local maxima are evaluated when multiple candidates are considered as described in the next section. The following steps would then be performed on each of those positions. The selected picture is taken from the reference sequence and the corresponding scaling is applied to the test image. Using a straight-forward 2D phase correlation the spatial offsets in both directions are retrieved. Finally, the MSE value can be calculated.

In [6] the authors use an additional highpass filter after the first DFT but in our case this led to worse results. Additionally we are using a different logarithm base for the scale calculation of $b_x = b_y = 1.029$. In anamorphic scaling often a stretch factor of $b_x^{10} = 1.33$ occurs.

III. Multiple Candidates

Many matching algorithms provide additional information besides the corresponding matching frame. Most of those described in [4] estimate the offset in both spatial directions as well. The 3D approach from the last section even estimates the scale in both directions. Sometimes the criterion for the best matching frame is misleaded because of those additional degrees of freedom.

In order to improve the match, multiple candidates are calculated. This can be easily implemented in all matching algorithms. Then, the most often occuring combination of spatial offsets in the best match is searched and all those matches which have different offsets are eliminated.

An example output of a typical temporal matching situation is shown in Figure 2. The estimation seems correct up to frame 68. Then, suddenly, frame 11 becomes the candidate for the next distorted frame. Later, on frame 71, the estimation continues in a correct manner. This mismatch has two disadvantages for quality estimation algorithms, especially those which model the human visual system (HVS) like e.g. PEVQ which is developed by OPTICOM based on [7]. The first drawback is, that two completely different frames are compared. However, the second one is more important: The quality model detects two temporal discontinuities which are not really present in the video sequence. This leads to a much lower quality score than the human observer would choose. The following algorithm eliminates this error.

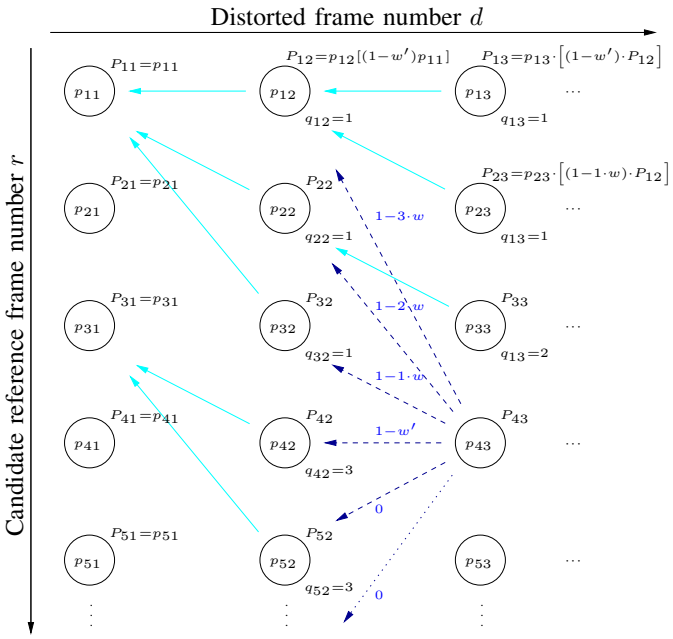

Fig. 3. Upper left section of the probability matrix

IV. Temporal smoothness constraint algorithm

In the following we denote the number of all reference frames by N_r and the number of all distorted frames by N_d. For this algorithm, we assume that there are several candidate reference frames which may match one distorted frame. In general, each distorted frame may be associated to one of all the reference frames available. Some of those associations are more likely than others which may be expressed by a probability $p(r,d)$. This leads to a probability matrix of size N_rxN_d. The upper left corner of that matrix is depicted in Figure 3. The result of the algorithm shall be a path in horizontal direction, selecting one candidate reference frame for each distorted frame from $d = 1$ to $d = N_d$. Without applying any restrictions, there are $N_r^{N_d}$ possible paths. Each leads to a unique conditional probability. However, the most probable path is requested. Therefore, a successive elimination algorithm can be applied. It is based on the idea that only the most probable path has to be stored and once it is found it does not change for further distorted frames.

This is explained with the example depicted in Figure 3. The calculation starts at position $(1,2)$. Assuming that even a distorted video sequence is causal, e.g. does not rewind to previous frames, only one predecessor is possible: $(1,1)$. Therefore the predecessor frame number 1 is stored in q_{12} and the conditional probability is calculated with a path malus w' as $P_{12} = p_{12}\cdot[(1-w')\cdot p_{11}]$ because the two selection events are statistically independent. In the next step, the most likely path to element $(2,2)$ is calculated. There are two possibilities, either the predecessor is $(1,1)$ or it is $(2,1)$. The one which leads to the highest probability is stored. In this example it is $(1,1)$. Because this leads to the highest probability, there will never be a path which leads to node $(2,2)$ and then to node $(2,1)$ because it would always be less suitable. Thus, the algorithm eliminates the second path, reducing the order of the problem to $O(N_r \cdot N_d)$. Although the complete path has to be estimated, it is sufficient to store only the predecessor reducing the amount of memory spent. The algorithm traverses first the columns and then the rows. Once it has reached the last distorted frame, the last column contains the joint probabilities of each most likely path which ends in that specific last frame. Therefore by finding the maximum in the last column and collecting the predecessors backwards along that selected path, the result is obtained.

The example shown in Figure 3 details the situation for the node $(4,3)$. In our algorithm, long jumps between frames shall become unlikely. Therefore we introduced a malus system for the path weights. The larger the distance to the previous frame the smaller the path weight gets. A linear weight with $w = 0.015$ is used in the evaluation. The situation concerning a temporal stay is different. As frame repeats are usually handled outside this algorithm, there is a larger malus for a temporal stay by using a special weight $w' = 0.15$. Therefore this is as expensive as a jump of ten frames. The probabilities are

TABLE I
Display distortions

Nr	Class	Type of distortion
1	A	No additional display distortion
	B	Spatial offsets
2		Spatial offset by (2,2) pixels
3		Spatial offset by (8,8) pixels
	C	Intensity distortions
4		Gamma correction using factor 1.4 on RGB
5		Contrast enhancement, factor 1.5
6		Brightness offset, 8% on Y-component
	D	Geometric distortions
7		Zoom by 16 pixels
8		Stretch by 4/3, e.g. aspect ratio correction
	E	Frequency located transmission distortions
9		Additive White Gaussian Noise on Y-component
10		Ringing on Y-component
11		Low-pass filter with cut-off frequency 0.5
	F	Colorspace transformations
12		256 Indexed colors, optimum palette
13		64k Indexed colors, fixed table RGB565
14		Histogram equalization

calculated from the MSE values of the different algorithms by first calculating $p'(r,d)$ for all those values which are available according to

$$p'(r,d) = \frac{1}{\sqrt{\mathrm{MSE}(r,d)}}$$

In our evaluation we have ten distinct MSE values per column. All other values are then set to half the minimum occuring in the complete matrix. Finally, the values are normed to a sum of one by

$$p(r,d) = \frac{p'(r,d)}{\sum_{\rho=1}^{N_r}\sum_{\delta=1}^{N_d} p'(\rho,\delta)}$$

Due to the small probabilities which result from the conditional probabilities it has proven necessary to do all calculations in logarithmic space.

V. Results

The evaluation results are presented similar to [4]. The same 12 sequences, display distortions and classes are used. The distortions are shown in Table I. As the VQEG group specified the temporal matching criteria in the meantime, a compatible temporal search range of $-0.6s \ldots + 2.6s$ was used and all 220 frames are considered. The proposed PC3D-SAD algorithm is compared to the best algorithms that were evaluated in [4]. In Table II the percentage of false decisions is shown when there is no video codec involved. Table III summarizes the results for H.264 coded sequences at 128 kbit/s. The detection rate of the PC3D-SAD is very good even without the additional Maximum-Likelihood (ML) step. However, all algorithms are improved on average by adding the smoothness constraint.

TABLE II

Detection results with lossless video coder transmission

Algorithm	+ML	Avg	False detection percentage for distortion class					
			A	B	C	D	E	F
MSE	no	23.3	0.0	70.1	2.9	63.5	0.0	3.6
PC-SAD	no	15.3	0.0	18.0	2.9	67.6	0.0	3.6
PC3D-SAD	no	3.5	0.0	2.9	2.4	12.6	0.0	3.3
MSE	yes	19.1	0.0	61.5	1.8	48.3	0.0	2.9
PC-SAD	yes	12.4	0.0	5.8	1.9	64.1	0.0	2.9
PC3D-SAD	yes	2.6	0.0	1.5	1.0	12.5	0.2	0.3

TABLE III

Detection results with additional distortion by H.264 coding

Algorithm	+ML	Avg	False detection percentage for distortion class					
			A	B	C	D	E	F
MSE	no	24.5	0.7	70.4	4.8	65.7	0.8	4.9
PC-SAD	no	17.9	0.7	27.0	4.8	69.4	0.8	5.0
PC3D-SAD	no	6.1	1.0	1.6	3.9	23.6	1.1	5.3
MSE	yes	20.1	0.3	63.9	2.2	50.7	0.3	3.5
PC-SAD	yes	15.1	0.3	18.1	3.1	65.3	0.3	3.8
PC3D-SAD	yes	3.9	1.2	0.4	2.1	17.0	1.4	1.6

Another interesting question would be how far the wrong matches are apart from the correct match. A squared difference measures seems appropriate, because a large deviation is much worse than a small one. We use the Root Mean Squared Error (RMSE) measure for each sequence:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_d} \cdot \sum_{t=1}^{N_d} (c_e - c_c)^2}$$

where c_e denotes the estimated corresponding reference frame and c_c would be the correctly matched frame. It is also equivalent to the standard deviation of the matching error if a mean value of 0 is assumed for the error. The results for all sequences are shown as an average in Table IV and in Table V. The tables show that the distance is often negligible for the PC3D-SAD with smoothness constraint.

VI. Conclusions

One of the most important steps in the evaluation of the video quality using a perceptual measure is the correct temporal and spatial alignment of the sequences. Only if this step is successful, all subsequent calculations will improve the correlation with subjective tests. The determination of the correct corresponding reference frame to a distorted frame is often very difficult because the distortions by the video coder or some postprocessing steps may be severe.

TABLE IV

RMSE distance results with lossless video coder transmission

Algorithm	+ML	Avg	Distance of mismatch measured in MSE for distortion class					
			A	B	C	D	E	F
MSE	no	2.66	0.00	6.24	0.63	7.07	0.00	2.00
PC-SAD	no	2.08	0.00	2.31	0.62	7.52	0.00	2.04
PC3D-SAD	no	0.74	0.00	1.67	0.37	0.91	0.02	1.49
MSE	yes	0.10	0.00	0.26	0.02	0.29	0.00	0.06
PC-SAD	yes	0.08	0.00	0.10	0.02	0.31	0.00	0.06
PC3D-SAD	yes	0.03	0.00	0.07	0.01	0.04	0.00	0.04

TABLE V

RMSE with additional distortion by H.264 coding at 128 kbit/s

Algorithm	+ML	Avg	Distance of mismatch measured in MSE for distortion class					
			A	B	C	D	E	F
MSE	no	2.66	0.03	5.79	1.25	7.00	0.04	1.87
PC-SAD	no	2.41	0.03	3.02	1.17	7.94	0.04	2.28
PC3D-SAD	no	1.63	0.61	0.59	0.99	4.65	0.49	2.47
MSE	yes	0.10	0.00	0.24	0.03	0.29	0.00	0.05
PC-SAD	yes	0.09	0.00	0.13	0.03	0.33	0.00	0.06
PC3D-SAD	yes	0.06	0.05	0.02	0.03	0.19	0.01	0.07

In this paper a combination of a matching algorithm and a maximum-likelihood algorithm was proposed. The matching algorithm is able to detect correspondences even if the image was shifted or scaled. Both are typical artifacts which often occur in transmission scenarios, especially if a resolution change or an aspect ratio adaptation is used. It has been shown that the new algorithm PC3D-SAD outperforms the previously examined algorithms in [4]. The maximum-likelihood algorithm puts an additional constraint on the resulting matches which does not allow rewinding the video sequence and puts a penalty on large skips.

The combination of these two algorithms results in a very stable detection. Less than four percent of the matches are wrong. Compared to our previous results in [4] this is an improvement by a factor of three.

References

[1] *Multimedia Group Test Plan*, Video Quality Experts Group, February 2007, draft Version 1.16.

[2] Q. Huynh-Thu and M. Ghanbari, "Impact of jitter and jerkiness on perceived video quality," in *Proc. of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, B. Li, Ed., Jan. 2006.

[3] M.Barkowsky, J.Bialkowski, and A.Kaup, "Subjective Video Quality Assessment for Low Bitrate Multimedia Applications (in german)," in *ITG Fachbericht 188: Elektronische Medien 2005*. VDE-Verlag, 2005, pp. 169–175.

[4] M. Barkowsky, R. Bitto, J. Bialkowski, and A. Kaup, "Comparison of matching strategies for temporal frame registration in the perceptual evaluation of video quality," in *Proc. of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, B. Li, Ed., Jan. 2006.

[5] J. G. Proakis, *Digital communications*, 3rd ed. McGraw-Hill Book Co., 1995.

[6] B. S. Reddy and B. Chatterji, "An FFT-Based Technique for Translation, Rotation, and Scale-Invariant Image Registration," in *IEEE Trans. on Image Processing*, Aug. 1996, vol. 5, no. 8, pp. 1266–1271.

[7] A. Hekstra, J. Beerends, *et al.*, "PVQM - A perceptual video quality measure," in *Signal Processing: Image Communications*, vol. 17. Elsevier, 2002, pp. 781–798.