CCITT SG XV
Working Party XV/1
Specialists Group on Coding for
Visual Telephony

Doc. No: 300
Jan. 88

Title: MOTION-COMPENSATING PREDICTION WITH FRACTIONAL-PEL-ACCURACY

Source: FRG

I. The problem

With motion-compensating prediction in a hybrid coder, the current picture s(x,y) is predicted by a signal $\hat{s}(x,y)$ that is obtained from a previously transmitted, reconstructed picture r(x,y) according to

$$\hat{s}(x,y) = f(x,y) ** r(x-\hat{d}_x, y-\hat{d}_y) \qquad (1)$$

In this equation $(\hat{d}_x, \hat{d}_y)$ is the estimated displacement vector, "**" denotes 2-D convolution and f(x,y) is the impulse response of the "filter-in-the-loop". f(x,y) has a twofold function:

- It is needed as interpolation filter to improve prediction by fractional-pel-accuracy of the motion compensation ("improved accuracy effect").

- It can improve prediction efficiency by taking into account the crosscorrelation between $\hat{s}(x,y)$ and $r(x-\hat{d}_x, y-\hat{d}_y)$ ("Wiener filter effect").

In order to achieve fractional-pel-accuracy, in many cases the "bilinear filter" with

$$f(x,y) = \max\{0,1-|x|\} \cdot \max\{0,1-|y|\} \qquad (2)$$

091/60

has been used. With the bilinear filter, we benefit from both the improved accuracy effect and the Wiener filter effect. It is not clear, however, how large contributions are from each individual effect.

In this study the following questions are addressed:

- With fractional-pel-accuracy of motion-compensation using a bilinear filter, how large are the contributions to improved prediction of the "improved-accuracy-effect" and of the "Wiener-filter-effect"?

- To what extent can we further improve prediction by filtering in the loop?

- How should the impulse response of the filter used for spatial interpolation in the context of fractional-pel-accuracy look like?

- What accuracy of displacement measurement and motion compensation makes sense?

The problems are investigated experimentally on the basis of video signals typical for videotelephone applications and for broadcast applications.

II. Theoretical background

It is shown in [1], that for a signal with exclusively translatorily moving picture contents the prediction error

$$e(x,y) = s(x,y) - \hat{s}(x,y) \tag{3}$$

possesses a spatial power spectral density

(4)

$$\Phi_{ee}(\Lambda) = \Phi_{ss}(\Lambda)\,[1+|F(\Lambda)|^2-2\mathrm{Re}\{F(\Lambda)P(\Lambda)\}]+\Phi_{nn}(\Lambda)|F(\Lambda)|^2$$

where
- Λ is the spatial frequency vector,
- $\Phi_{uu}(\Lambda)$ is the spatial power spectral density of a signal u
- $F(\Lambda)$ is the 2-D Fourier transform of f(x,y)
- $P(\Lambda)$ is the 2-D Fourier transform of the displacement estimation error p.d.f.
- $\mathrm{Re}\{\cdot\}$ denotes the real part of a complex number
- n(x,y) represents noise or other signal components, that are contained in r(x,y), but not in s(x,y)

The mean squared prediction error is minimized for each spatial frequency, if f(x,y) is a Wiener filter with the frequency response

$$F(\Lambda) = P^*(\Lambda)\,\frac{\Phi_{ss}(\Lambda)}{\Phi_{ss}(\Lambda)+\Phi_{nn}(\Lambda)} \qquad (5)$$

This Wiener filter can be interpreted to consist of two stages, one of which is a Wiener filter with respect to the noise n(x,y) contained in the reconstructed signal r(x,y), and the other one is a Wiener filter that takes into account the inaccuracy of the displacement estimate.

Model calculations allow to study the influence of displacement estimation error variance on the prediction error variance. Details of this analysis can be found in [1]. Fig. 1 shows the results of these model calculations for different levels of white noise n contained in the previous reconstructed picture r, namely for SNR = 10 dB, 20 dB and 30 dB. The cases "Wiener filter" according to (5) and "no filter", i.e.

$$F(\Lambda) = 1 \qquad (6)$$

are compared.

The following observations are important:

- Prediction error variance is generally decreased by more accurate motion compensation.
- Beyond a certain "critical accuracy" the possibility of further improving prediction by more accurate motion compensation is very small.
- The critical point is at a high displacement estimation error variance for low SNR and at low displacement estimation error variance for high SNR.
- For high SNR the Wiener filter is more effective for less accurate displacement estimation than for accurate displacement estimation.
- For low SNR the Wiener filter is more effective for accurate displacement estimation than for less accurate displacement estimation.
- For accurate displacement estimation the potential of the Wiener filter increases with noise level.

III. Displacement estimation with fractional-pel-accuracy

After considering the theory of section II, we will measure prediction error variance as a function of displacement vector accuracy with real pictures. In order to obtain meaningful results, it is very important to use a displacement estimator, that gives both reliable and accurate results.

We have used a most sophisticated displacement estimation algorithm, that consists of three stages:

A. Displacement measurement by phase correlation

Large overlapping measurement windows (64x64) are moved over the picture. For each measurement window, a limited number of candidate vectors is extracted with integer-pel-accuracy from a phase correlation surface [2].

B. Segmentation by candidate vector assignment

The candidate vector of the corresponding measurement window, that provides the best match for a smaller "block", is assigned to that block. The size of the (non-overlapping) blocks was 16x16 in our experiments.

C. Displacement vector refinement by search

For fractional-pel-accuracy a block matching search procedure starts with the vector from stage B. The search procedure successively considers displacements of 1/2 pel, 1/4 pel etc. until the desired accuracy has been obtained.

Stages A and B of this estimator have been proposed by BBC researchers recently [3]. With the appropiate parameter setting the estimated vector fields are both reliable and accurate. Vectors fields are smooth within homogenously moving regions, while corona effect around moving objects are avoided (Fig. 2). With respect to mean squared prediction error the estimator outperforms other state-of-the-art schemes like hierarchical blockmatching or differential techniques for a given vector accuracy and blocksize.

IV. An experimental comparison of different spatial interpolation filters

Three different types of spatial interpolation filters f(x,y) have been compared at different displacement estimation accuracies.

1. Sinc-interpolation corresponds to the case $F(\Lambda) = 1$ ("no filter" in Fig. 1). For sampled lowpass signals the corresponding impulse response ideally would have a sin(x)/x-shape. For a practical system, however, the infinite impulse response has to be truncated to a finite impulse response. For fractional-pel-accuracy we have cascaded horizontal and

vertical 2:1 interpolations with carefully designed 21-tap filters.

2. Bilinear interpolation uses the impulse response (2). For integer-pel-accuracy of the displacement vector sinc-interpolation and bilinear interpolation are identical.

3. Wiener filters have been computed for each picture pair separately. Eq. (5) gives the Wiener filter as a function of signal and noise power spectra and of displacement estimation error p.d.f.. Since noise power spectrum and displacement estimation error p.d.f. are not known explicitly, an alternative approach has to be used to determine the Wiener filter. It is easy to show [4], that the Wiener filter possesses the frequency response

$$F(\Lambda) = \frac{\Phi_{sc}(\Lambda)}{\Phi_{cc}(\Lambda)} \qquad (7)$$

where $\Phi_{sc}(\Lambda)$ - cross spectrum of the motion-compensated picture c(x,y) and the current picture s, with $c(x,y) = r(x-\hat{d}_x, y-\hat{d}_y)$

$\Phi_{cc}(\Lambda)$ - power spectrum of c(x,y)

With the assumption, that the previous reconstructed picture r(x,y) does not differ from the previous original picture, thus neglecting coding noise, $\Phi_{sc}(\Lambda)$ and $\Phi_{cc}(\Lambda)$ have been measured directly from the signal. For picture pairs TREVOR and MISS AMERICA only moving parts have been considered for the design of the Wiener filter.

The mean squared prediction error for the different spatial interpolation filters is shown in Fig. 3 and Fig. 4 as a function of displacement estimation accuracy, again neglecting coding noise in the previous picture r(x,y). For TREVOR and MISS AMERICA only moving parts were considered. Figs. 3 and 4 illustrate the following observations:

- Compared to integer-pel-accuracy of motion-compensation without filtering, prediction efficiency can be improved by more accurate motion-compensation and Wiener filtering by up to 5.2 dB for ZOOM, 2.3 dB for VOITURE, 1.8 dB for TREVOR and 0.7 dB for MISS AMERICA.
- Except for ZOOM, bilinear interpolation is as good or better than sinc-interpolation.
- For the broadcast TV signals ZOOM and VOITURE, 1/4-pel-accuracy in the motion-compensating predictor is certainly sufficiently accurate for a practical coder.
- For the videophone signals TREVOR and MISS AMERICA 1/2-pel-accuracy seems to be a desirable limit.
- The curves in Fig. 3 and Fig. 4 qualitatively correspond to the model curves in Fig. 1. The measurements can be explained by the theory presented in section II.

For the videophone signals and for 1/2-pel-accuracy of motion-compensation, we have simplified the Wiener filter to a separable filter that is suitable for a hardware realization. This filter uses the coefficients (1/8,6/8,1/8) for samples centered around the estimated motion trajectory, if the estimated displacement vector component is integer. If the estimated displacement requires a half-pel-shift, the coefficients (1/16,7/16,7/16,1/16) are used. This filter, denoted by (1,2,7,12,7,2,1), is compared in Table 1 with bilinear interpolation and with Wiener filtering. The additional gains over bilinear interpolation are fairly small.

V. Conclusion

We have studied the effects of fractional-pel-accuracy of displacement estimation on the efficiency of a motion-compensating predictor in conjunction with different spatial interpolation filters.

A three-stage procedure for reliable displacement estimation with fractional-pel-accuracy has been described.

A "critical accuracy", beyond which the possibility of further improving prediction by more accurate motion compensation is small, is due to noise components in the previous reconstructed picture, that do not obey the paradigm of translatory motion. For typical broadcast TV signals 1/4-pel-accuracy in the motion-compensating predictor appears to be sufficient, while for videophone signals 1/2-pel-accuracy is desirable.

For videophone signals with bilinear interpolation the "Wiener filter effect" is partly exploited additionally to the "improved accuracy effect". Another separable filter, that gives slightly better prediction, has been proposed.

VI. References

[1] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," IEEE Journal on Selected Areas in Communications, vol. SAC-5, no. 7, 1140 - 1154, August 1987

[2] C.D. Kuglin, D.C. Hines, "The phase correlation image allignment method," Proc. of the IEEE 1975 Intern. Conf. on Cybernetics and Society, 163 - 165 September 1975.

[3] G.A. Thomas, "Television motion measurement for DATV and other applications," BBC Research Department Report no. 1987/11, September 1987.

[4] A. Papoulis, "Probability, Random Variables, and Stochastic Processes," McGraw-Hill, 1965.

Appendix: Picture material used for the experiments

We have selected four picture pairs for our experimental investigations.

ZOOM: Two fields out of a broadcast TV sequence (DOLL40 kindly provided by Deutsche Thomson-Brandt), sampled at 13.5 MHz with line-interlace. The two fields are taken 40 ms apart. The window, that has been processed, is 128 lines x 256 pels and shows a building and parts of a ship with a lot of detail. The movement is generated exclusively by camera zoom.

VOITURE: Two fields 20 ms apart from a broadcast TV sequence (VOITURE kindly provided by CCETT) sampled at 13.5 MHz with line-interlace. Mainly horizontal motion of rigid objects due to camera pan, motion of the car and motion of the gate. Contains a lot of decovered background. Processed window: 256 lines x 512 pels.

TREVOR: A videophone sequence (TREVOR kindly provided by British Telecom Research Labs) has been converted to a format of 7.5 non-interlaced fields per seconds with 288 lines x 352 pels. The fields no. 10 and 11 of this sequence have been processed. The processed window is 256 lines x 256 pels. The picture pair represents considerable motion (up to 9 lines vertically and up to 6 pels horizontally) combined with significant object deformation.

MISS AMERICA: A videophone sequence which again has been converted to a format of 7.5 non-interlaced fields per seconds with 288 lines x 352 pels. The fields no. 5 and 6 of this sequence have been processed. The processed window is 256 lines x 256 pels. The picture pair represents moderate motion combined with a significant change in facial expression.

891

Table 1 - Mean squared prediction error for moving areas, comparison of interpolation with the filter (1,2,7,12,7,2,1) (see text) with bilinear interpolation and with Wiener filtering. Motion-compensation is done with 1/2-pel-accuracy.

signal	bilinear interpolation	filter (1,2,7,12,7,2,1)	Wiener filter
TREVOR	97.6	92.4	89.5
MISS AMERICA	147.5	146.5	142.5

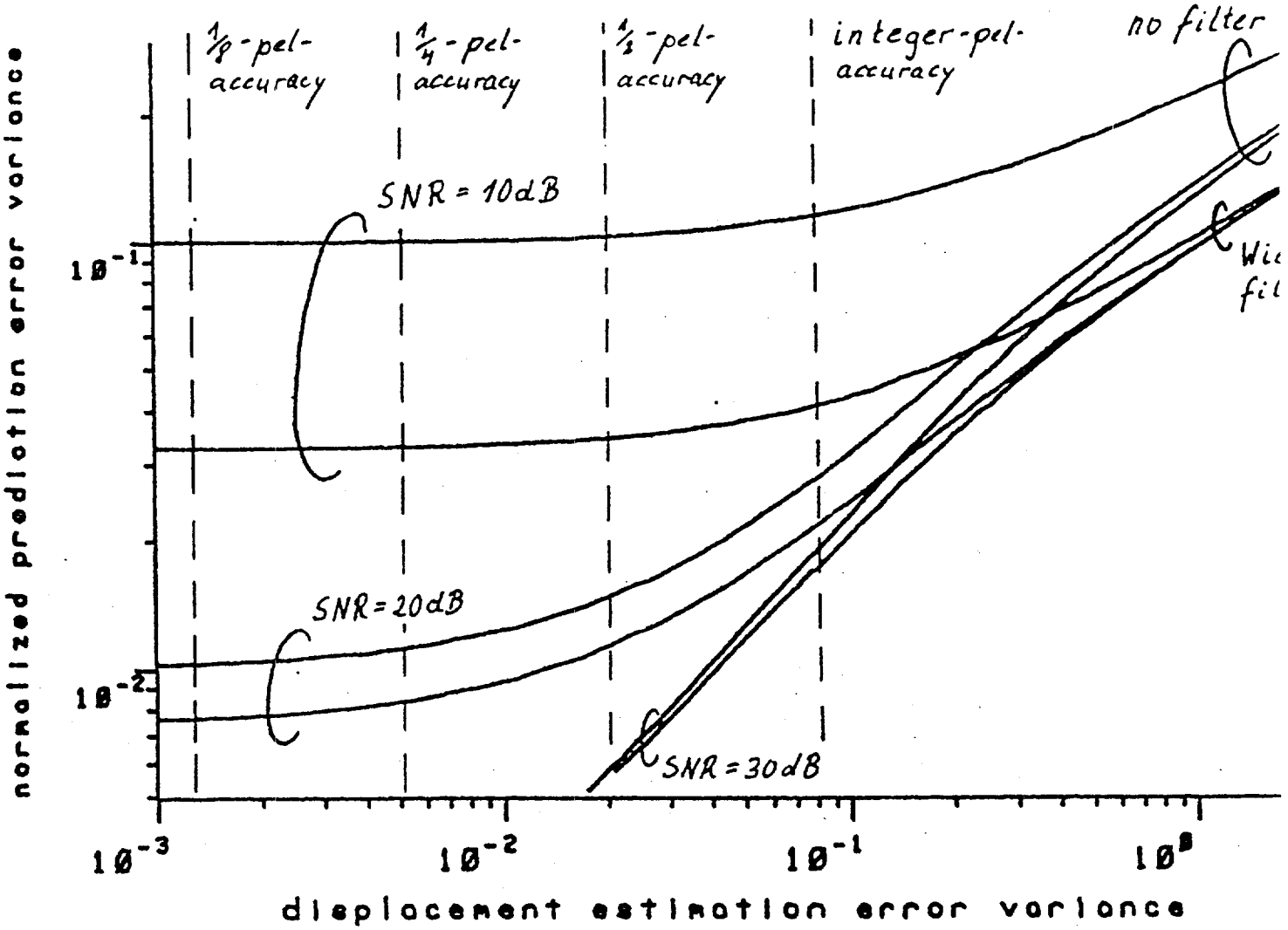


Fig. 1 - Results of a model calculation showing the influence of displacement estimation accuracy on prediction efficiency for noisy signals. The vertical dashed lines indicate the minimum displacement estimation error variances, that can be achieved with the indicated displacement vector accuracy in moving areas.

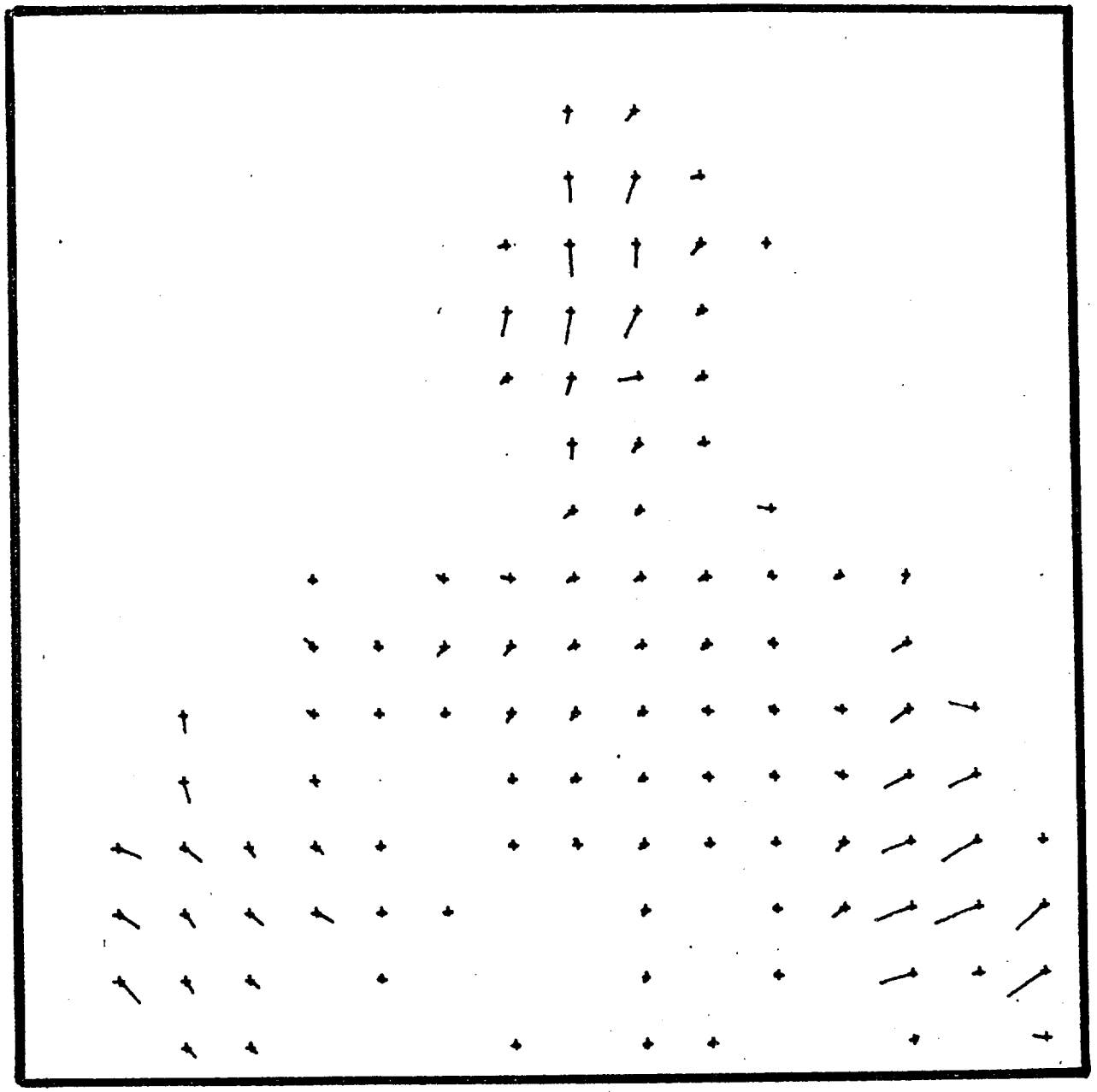


Fig. 2 - Example of a displacement vector field out of a well-known videophone testsequence estimated by the three-stage-procedure described in the text

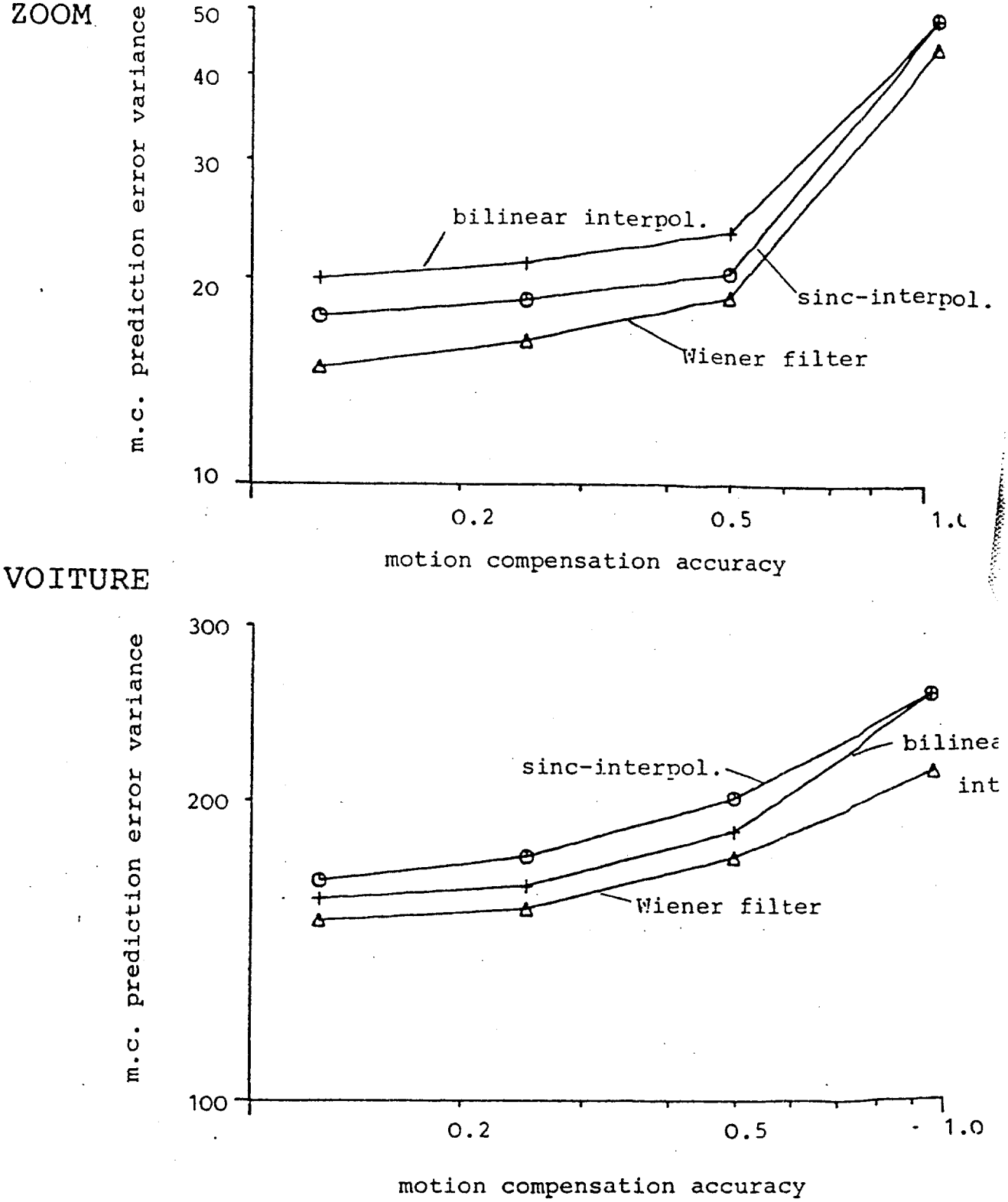
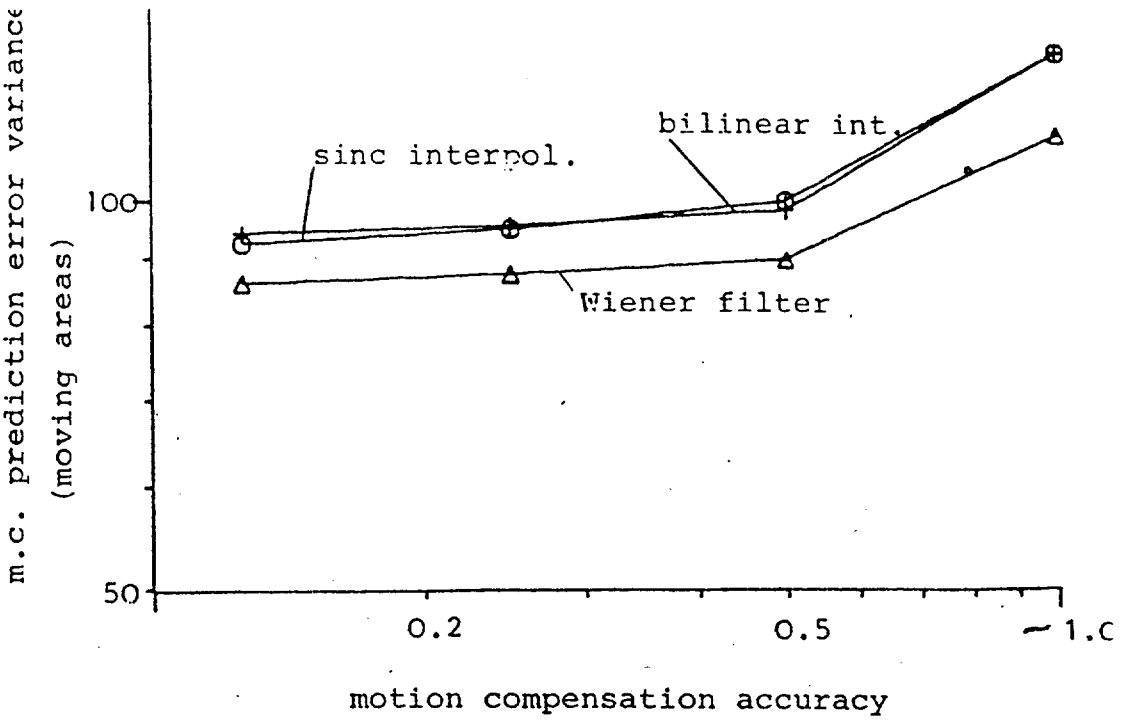


Fig. 3 - Experimental comparison of different spatial interpolation/ prediction filters for broadcast TV signals

TREVOR



MISS AMERICA

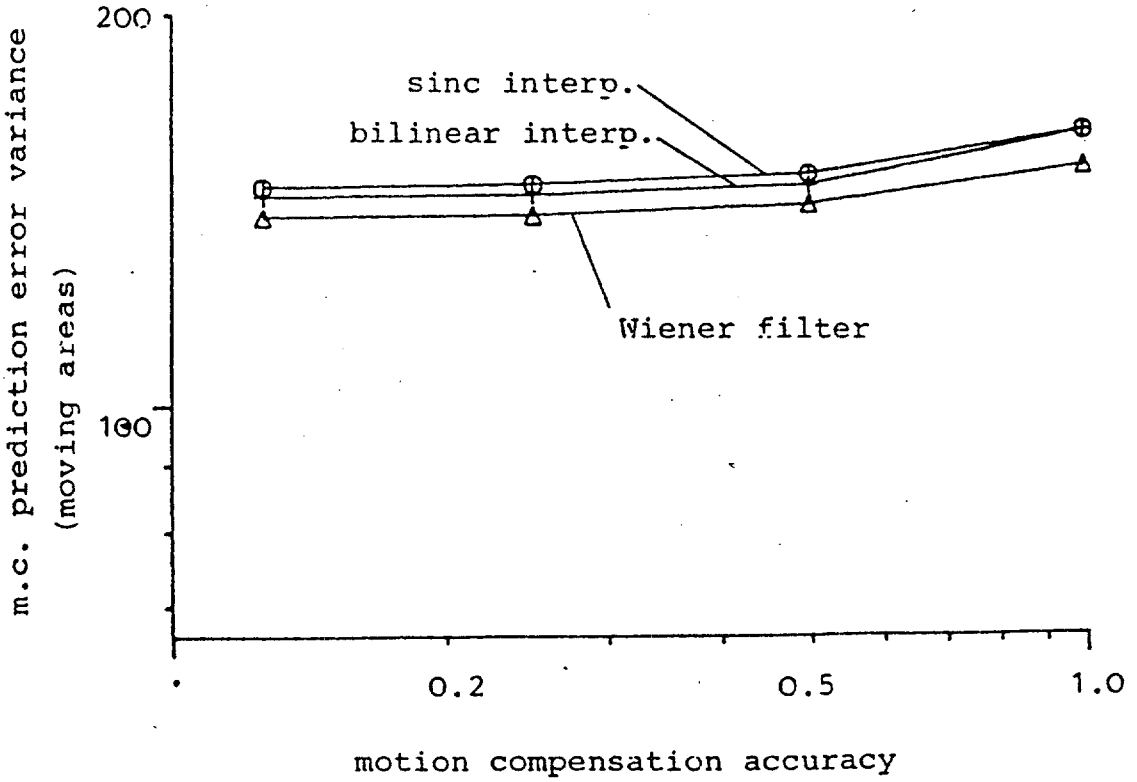


Fig. 4 - Experimental comparison of different spatial interpolation/ prediction filters for videophone signals