SPATIALLY ADAPTIVE WAVELET TRANSFORM FOR VIDEO CODING WITH MULTI-SCALE MOTION COMPENSATION

Marta Mrak and Ebroul Izquierdo

Multimedia and Vision Research Group, Queen Mary, University of London
Mile End Road, E1 4NS, London, UK
{marta.mrak, ebroul.izquierdo}@elec.qmul.ac.uk

ABSTRACT

In this paper a technique that enables efficient synthesis of the prediction signal for application in multi-scale motion compensation is presented. The technique targets prediction of high-pass spatial subbands for motion compensation at higher scales. Since in the targeted framework these subbands are obtained by high-pass filtering of prediction signal in pixel domain, an adaptive approach for filtering is proposed to support decomposition of differently predicted frame areas. In this way an efficient application of different prediction modes at all scales used for compensation is enabled. Experimental results show that for fast sequences where such an application of different prediction modes is crucial, the proposed adaptive transform introduces significant objective and visual improvements.

Index Terms — adaptive transforms, multi-scale compensation, scalable video coding

1. INTRODUCTION

Motion compensation in conventional video coding is the main contributing factor in the compression performance. However, for its application in scalable video coding specific constraints have to be taken into account so that the targeted scalability functionalities can be achieved. In wavelet-based video coding the temporal and spatial decompositions support temporal and spatial scalabilities. Flexibility of the wavelet-based scalable coding systems emerges from possibility to apply those decompositions in different orders so that performance in different spatio-temporal decoding points is optimised. However, since the decomposition steps are interleaved and non-commutative, the design of temporal decomposition modules has to take into account requirements for spatial scalability. Those refer to improved decoding performance at lower spatial resolutions which can be achieved using multi-scale motion compensation.

A technique that improves prediction for decomposition schemes that use motion compensation at multiple scales is proposed. In these schemes the prediction signal is created in frame domain with actual compensation at the level of low- or high-pass spatial subbands. The efficiency of temporal decomposition in such schemes depends on a number of factors related to the applied motion model, motion estimation and temporal filtering. Common approach for achieving good compensation is to use adaptive prediction mode selection for different frame areas. Since those structurally different areas have to be spatially transformed, application of wavelet transform on the whole predicted frame may cause unacceptable artefact at the discontinuities between differently compensated areas when decoding at low bit-rates. Therefore we propose a strategy to adapt wavelet transform in order to achieve better prediction signal representation.

In the past several approaches for spatially adaptive transform have been proposed for application in video coding. In [1] wavelet transform is adapted according to the content of the original or compensated frame content. Since it does not take into account the prediction type of specific frame areas it cannot be efficiently applied in the given scenario. On the other hand, in [2] the compensated frames are spatially transformed taking into account the structure given by motion compensation such that the adaptation of wavelet transform is here driven by motion information. While in that work the goal was to achieve better energy compaction, leading to better compression, in this paper the idea of adapting wavelet transform according to motion information is extended to the application of better prediction signal synthesis for motion compensation at higher scales in schemes with compensation at multiple scales.

Section 2 introduces the concept of multi-scale motion compensation. The basic motion-driven spatial transform with application in such schemes is explained in Section 3 where a new application of such transform is also proposed. The comparative results of encoding with and without the newly introduced approach are summarised in Section 4, while Section 5 concludes this paper.

2. MULTI-SCALE MOTION COMPENSATION

In wavelet-based scalable video systems the conventional application of temporal decomposition followed by spatial decomposition does not yield acceptable results when lower spatial resolutions are targeted at the decoder. This problem has been resolved by application of temporal decomposition, i.e. motion compensation, at several spatial resolution levels [3-7]. While such approaches enhance the performance at lower resolutions, for higher resolutions the compression performance will be degraded due to poorer compensation in spatial high-pass subbands or due to insufficient prediction between spatial subbands originating from different resolutions.

An efficient technique that gives good performance at all targeted resolution has been proposed in [6]. It combines motion estimation and compensation at each spatial resolution and is non-expansive, i.e. it keeps the resulting number of transformed coefficients equal to the number of input pixels. Therefore such

This research was partly supported by the European Commission under contract FP6-001765 aceMedia.

solution that uses multi-scale compensation can be also used in generalised spatio-temporal decompositions [8] where an arbitrary decomposition order can be selected, providing flexible adaptation to targeted spatio-temporal decoding points. An example that illustrates multi-scale compensation of frames at two spatial levels is shown in Figure 1. In the notation for frames $\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}$ a subscript t, L denotes non-compensated subbands (L) of t-th temporal level, and superscript s, L denotes frames (L) at s-th spatial level. Compensated frames (high-pass temporal subbands) and high-pass spatial subbands are represented with H. At both scales the motion warping $\mathcal{W}$, i.e. synthesis of the prediction signal, is performed on frame pixels. At the lowest spatial resolution ($s = S$) the compensation is, in this case of dyadic decomposition, performed on even frames as:

$$\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{L}}\left(k\right) = K_H \cdot \left(\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}\left(2\cdot k+1\right) - \mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}\left(n\right)\Big|_{\forall n\in\mathrm{P}}\right\}\right)\right),$$

where K_H is the normalisation factor for a specific decomposition scheme and the prediction signal is synthesised from frames with indices from set P. For the examples in Figure 1 this set is defined as P = $\{2 \cdot k, 2 \cdot k + 2\}$. The index of compensated frame $\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{L}}\left(k\right)$ originates from applied dyadic decomposition ($2\,k + 1 \rightarrow k$).

For multi-scale compensation, at higher spatial resolutions ($s > S$) the compensation is performed on high-pass spatial subbands as:

$$\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{H}}\left(k\right) = K_H \cdot \left(\mathcal{A}_{\mathrm{S}}^{\mathrm{H}}\left(\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}\left(2\cdot k+1\right) - \mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}\left(n\right)\Big|_{\forall n\in\mathrm{P}}\right\}\right)\right)\right), \quad (1)$$

where the set P is here defined in the same way as for resolution s and $\mathcal{A}_{\mathrm{S}}^{\mathrm{H}}$ denotes spatial decomposition with high-pass subband selection. With this notation, spatial decomposition $\mathcal{A}_{\mathrm{S}}$ of a frame can be expressed as:

$$\mathcal{A}_{\mathrm{S}}\left(\mathbf{f}^{s,\mathrm{L}}\right) = \left\{\mathbf{f}^{s-1,\mathrm{L}}, \mathbf{f}^{s,\mathrm{H}}\right\} = \left\{\mathcal{A}_{\mathrm{S}}^{\mathrm{L}}\left(\mathbf{f}^{s,\mathrm{L}}\right), \mathcal{A}_{\mathrm{S}}^{\mathrm{H}}\left(\mathbf{f}^{s,\mathrm{L}}\right)\right\}.$$

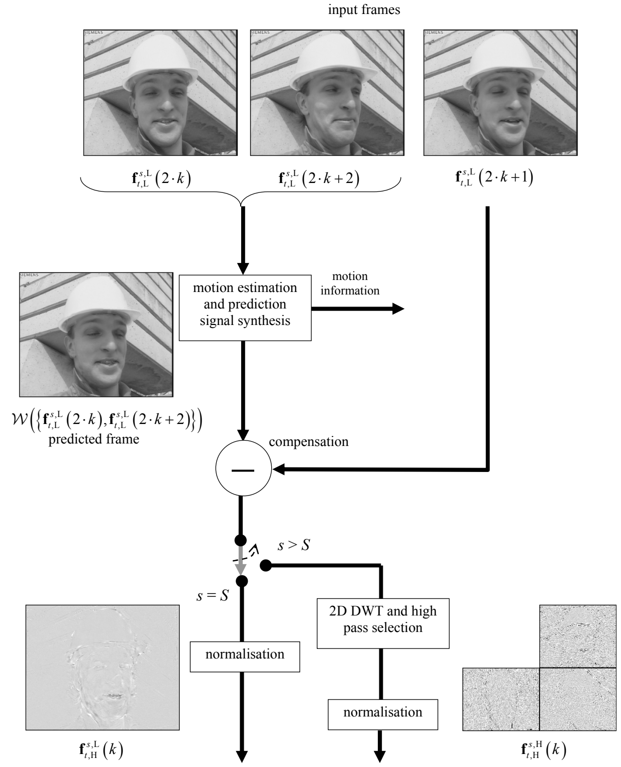

input frames
$\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k)$
$\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k+2)$
$\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k+1)$
motion estimation and prediction signal synthesis
motion information
$\mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k), \mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k+2)\right\}\right)$
predicted frame
compensation
$s > S$
$s = S$
2D DWT and high pass selection
normalisation
normalisation
$\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{L}}(k)$
$\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{H}}(k)$

Figure 1 Frame prediction and compensation in video coding with multi-scale compensation.

Since at higher resolutions the spatial transform is applied to the compensated frame in order to obtain high-pass spatial subbands only, the energy in those subbands directly depends on the chosen transform. Because of the reversibility of such decomposition the applied filterbanks have to be the same as used in the downsampling process for obtaining sequence at lower scales e.g. biorthogonal 9/7. However, as shown in the following section, spatial decomposition can be separately performed on the prediction signal and the current frame, allowing for application of different decomposition strategies and leading to better representation of final compensated high-pass spatial subbands.

3. ADAPTIVE TRANSFORM IN MULTI-SCALE MOTION COMPENSATION

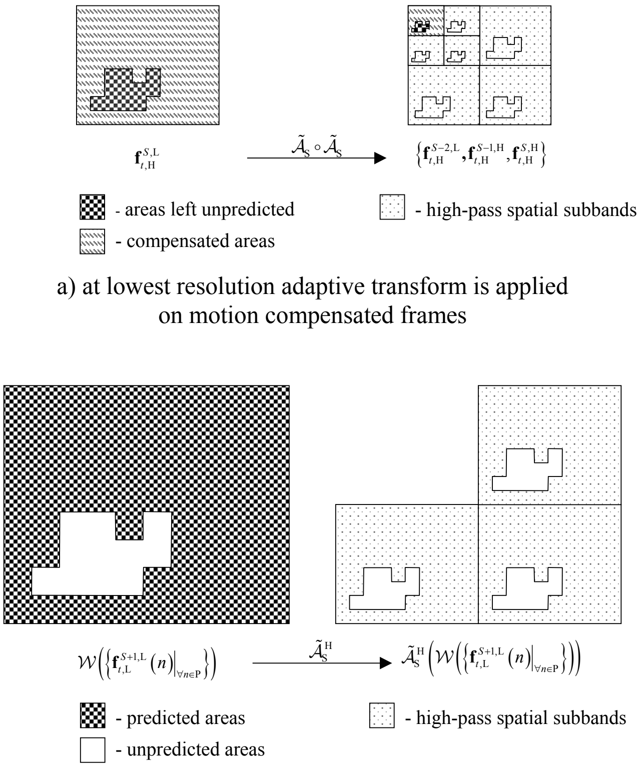
Recently a way to adapt spatial transform according to motion information has been developed, [2], with a goal to enable better signal representation and thus better compression and improved decoding quality. This so-called Motion-Driven Adaptive Transform (MDAT) is especially beneficial when intra blocks exist in high-pass frames. It has originally been proposed for application in schemes that use motion compensation at single scale. Here it is extended for application in schemes that use multi-scale motion compensation.

In the originally proposed format, MDAT has been used to establish a connection between motion-dependant high-pass frame content and further spatial decomposition. Adaptation in that way takes into account motion information and is realised using a separable lifting implementation of wavelet transform, [9]. For each lifting step l, applied on signal a_0 which corresponds to a row or a column of a frame or its low-pass representation, the adaptation is achieved by adaptation factors α_l and β_l:

$$a_l\left(k\right) = a_{l-1}\left(k\right) + \lambda_l\left(\alpha_l\left(k\right)\cdot a_{l-1}\left(k-1\right) + \beta_l\left(k\right)\cdot a_{l-1}\left(k+1\right)\right), (2).$$

While λ_l are the lifting coefficients that correspond to a particular wavelet filterbank, adaptation factors are defined by motion information. For non-adaptive transform $\alpha_l(i) = \beta_l(i) = 1$. If neighbouring pixels or transform coefficients $a_l(k)$ and $a_l(k + 1)$ belong to differently compensated areas (intra or inter), adaptation of transform for $a_l(k)$ is realised by choosing $\alpha_l(k) = 2$ and $\beta_l(k) = 0$ for each l, and $\alpha_l(k + 1) = 0$ and $\beta_l(k + 1) = 1$ for adaptation on $a(k + 1)$. Adaptation can be used at all levels of spatial transform, e.g. as suggested in Figure 2.a) where it is used twice. Adaptive transform is denoted as $\tilde{\mathcal{A}}_{\mathrm{S}}$. In this example the actual adaptation would be performed between compensated areas and areas left unpredicted. With this approach these structurally different areas of compensated frame are treated separately. Ringing artefacts that would occur on the border between different areas, and which can be regarded as sharp edge, are therefore avoided.

This approach can be applied at the lowest scale in schemes with compensation at multiple scales. Since at higher scales the compensation is performed only on high-pass spatial subbands, there is commonly no need for spatial transform of resulting compensated subbands.


$\mathbf{f}_{t,\mathrm{H}}^{S,\mathrm{L}}$
$\tilde{\mathcal{A}}_{\mathrm{S}} \circ \tilde{\mathcal{A}}_{\mathrm{S}}$
$\left\{\mathbf{f}_{t,\mathrm{H}}^{S-2,\mathrm{L}}, \mathbf{f}_{t,\mathrm{H}}^{S-1,\mathrm{H}}, \mathbf{f}_{t,\mathrm{H}}^{S,\mathrm{H}}\right\}$
▩ - areas left unpredicted
⬚ - high-pass spatial subbands
▤ - compensated areas

a) at lowest resolution adaptive transform is applied on motion compensated frames

$\mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{S+1,\mathrm{L}}(n)\Big|_{\forall n\in\mathrm{P}}\right\}\right)$
$\tilde{\mathcal{A}}_{\mathrm{S}}^{\mathrm{H}}$
$\tilde{\mathcal{A}}_{\mathrm{S}}^{\mathrm{H}}\left(\mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{S+1,\mathrm{L}}(n)\Big|_{\forall n\in\mathrm{P}}\right\}\right)\right)$
▩ - predicted areas
⬚ - high-pass spatial subbands
□ - unpredicted areas

b) at higher resolutions the adaptation is used on prediction signal

Figure 2 Application of adaptive transform in different stages of multi-scale compensation scheme.

In schemes with compensation at multiple scales the spatial transform firstly has to be performed on original frames in order to obtain sequence at lower resolution. Secondly, at lowest scale, after application of required number of temporal decomposition levels, spatial transform is used to further decompose high- and low-pass temporal subbands. Additionally, spatial transform is applied for the compensated frames at higher scales. (1) can be broken into two separate terms as:

$$\mathbf{f}_{t,\mathrm{H}}^{s,\mathrm{H}}(k) = K_H \cdot \left(\mathcal{A}_{\mathrm{S}}^{\mathrm{H}}\left(\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(2\cdot k+1)\right) - \tilde{\mathcal{A}}_{\mathrm{S}}^{\mathrm{H}}\left(\mathcal{W}\left(\left\{\mathbf{f}_{t,\mathrm{L}}^{s,\mathrm{L}}(n)\Big|_{\forall n\in\mathrm{P}}\right\}\right)\right)\right),$$

i.e. spatial decomposition can be performed separately on the current frame (or can be obtained from spatial decomposition resulting from downsampling) and on the prediction signal. If the prediction signal is a close match to the current frame that has to be compensated, then the common approach is to apply the same non-adaptive wavelet filterbank to both signals. However, if the prediction signal is not a close match to the current frame, an adaptation strategy can be used. Hence the corresponding spatial transform is marked as $\tilde{\mathcal{A}}_{\mathrm{S}}^{\mathrm{H}}$. Since the synthesis of the prediction signal in the spatial domain is driven by motion information, the prediction signal is structurally motion dependent. Therefore the application of motion adaptive high-pass filtering of prediction signal can be beneficial for obtaining better prediction of those areas that need to be compensated. In this case, it is required that unpredicted areas in the final compensated frame remain the same as in the high-pass subband of the current frame.

Relations that drive the adaptation using lifting in this case are identical to those from (2) with the difference that the actual adaptive transform is applied to the prediction signal (not to the compensated frame) and for only one level of spatial decomposition, as shown in Figure 2.b).

4. EXPERIMENTAL RESULTS

The tests have been performed in the video coding scheme with multi-scale motion compensation implemented in [7]. The chosen test material is a sequence with fast motion and many occluded areas - test sequence "Soccer", CIF (352 × 288) resolution, 15 fps. Such characteristics are common for real-life videos and are difficult for compression because of large number of areas that cannot be efficiently compensated. These areas are in motion estimation process marked as intra coded areas and in this experiment the actual compensation has been performed using the traditional fixed wavelet transform and the proposed spatially adaptive wavelet transform at all scales. The encoding configuration targets decoding at CIF and QCIF (176 × 144) resolutions.

One encoded sequence has been decoded at multiple bit-rates at original resolution for which the proposed adaptive transform is designed. Adaptation is performed at all colour components and the obtained PSNR results are averaged over the whole sequence. Since the proposed approach is applied only at higher spatial scales, the results at lower scales are not changed. Since fixed transform cannot efficiently represent prediction for high-pass spatial subbands, the proposed approach using adaptive transform achieves a significant gain at higher scale, as can be seen from Figure 3 where the PSNR results are presented for all colour components. In Figure 4 one frame of the decoded sequence is displayed, showing the visual difference between the applications of the adaptive compared to the fixed transform. The amount of visual artefacts depends on the amount of intra blocks, which for this particular example of fast moving scene is high. It can be seen that application of adaptive transform on the prediction signal results in better visual quality. That is because the ringing artefacts that would normally occur between differently encoded areas are avoided by the adaptation of spatial transform.

5. CONCLUSIONS

In this paper a new application of motion driven adaptive transform is proposed for use in video coding with motion compensation at multiple scales. In this case, the adaptive transform is used in motion compensation at higher scale for high-pass spatial subbands. More specifically, the presented approach enables better prediction in cases where parts of frame cannot be efficiently predicted and are therefore selected as intra coded areas.

For systems that use fast motion estimation algorithms and in which accurate prediction is not possible, this solution is beneficial as there are many frame areas that cannot be predicted from neighbouring frames. On the other hand if a sequence can be efficiently predicted, the adaptive transform does not have influence on the overall compression since the adaptation is only performed on these specific areas.

The introduced computational complexity of the proposed method is low as only a few additional multiplications are introduced per lifting step in wavelet transform. Since in this case the same motion information is available at both encoder and decoder, transmission of additional side information is not needed.

As a consequence of better motion compensation resulted from application of adaptive transform, the proposed scheme introduces considerable gain when applied in scalable video coding.

6. REFERENCES

[1] G.C.K. Abhayaratne, "2D wavelet transforms with a spatially adaptive 2D low pass filter," IEEE Nordic Signal Processing Conference (NORSIG) 2004, pp. 93-96, June 2004.

[2] N. Sprljan, M. Mrak, E. Izquierdo, "Motion driven adaptive transform based on wavelet transform for enhanced video coding," Proc. 2nd International Mobile Multimedia Communications Conference (Mobimedia 2006), September 2006.

[3] N. Adami, M. Brescianini, M. Dalai, R. Leonardi, A. Signoroni, "A fully scalable video coder with inter-scale wavelet prediction and morphological coding," in Proc. SPIE Visual Communications and Image Processing, vol. 5960, (Beijing, China), July 2005.

[4] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis and P. Schelkens, "In-band motion compensated temporal filtering," Signal Processing: Image Communication, Vol. 19, No. 7, pp. 653 - 673, August 2004.

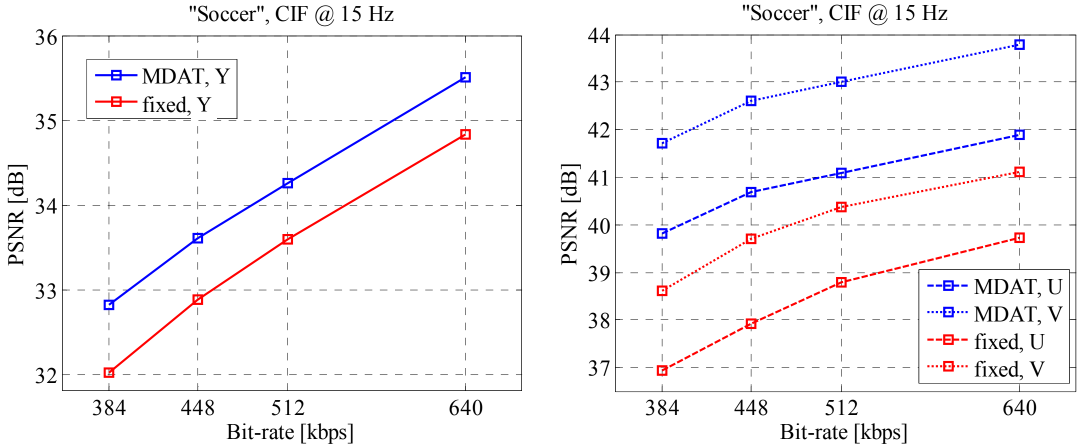
[5] N. Mehrseresht, D. Taubman, "A flexible structure for fully scalable motion compensated 3D-DWT with emphasis on the impact of spatial scalability," IEEE Trans. Image Processing, vol. 15, pp. 740 - 753, March 2006.

[6] R. Xiong, J. Xu, F. Wu, S. Li, "In-scale motion aligned temporal filtering," Proc. IEEE International Symposium on Circuits and Systems (ISCAS 2006), pp 3017-3020, May 2006.

[7] N. Sprljan, M. Mrak, T. Zgaljic, E. Izquierdo, *Software proposal for Wavelet Video Coding Exploration group*, ISO/IEC JTC1/SC29/ WG11/MPEG2005, no. M12941, 75th MPEG Meeting, January 2006.

[8] C. Ong, S. Shen, M. Lee, Y. Honda, "Wavelet video coding - generalized spatial temporal scalability (GSTS)," MPEG, ISO/IEC JTC1/SC29 WG11, M11952, April 2005.

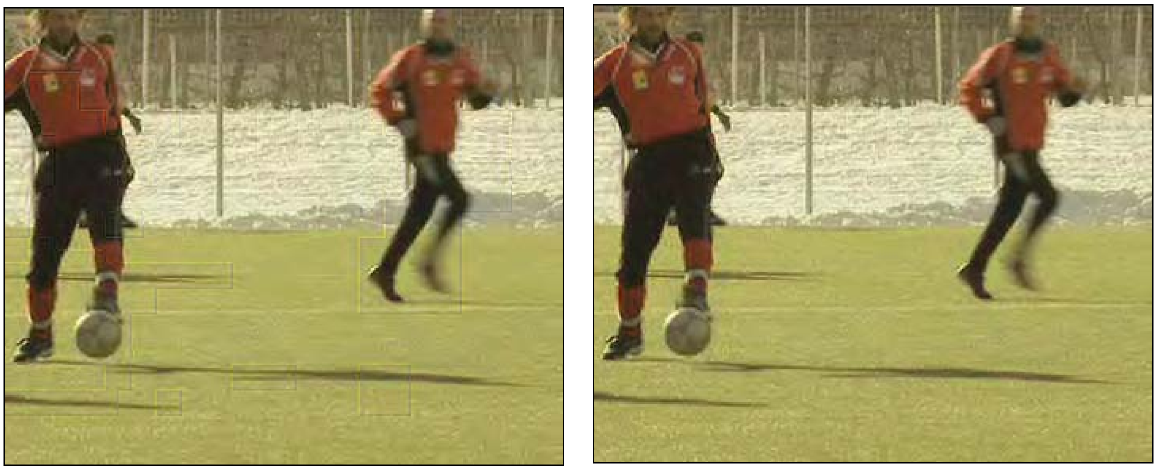
[9] Daubechies, I. and Sweldens, W., "Factoring Wavelet Transforms into Lifting Steps", J. Fourier Anal. Appl., Vol. 4, Nr. 3, 1998.



a) results for the luminance component

b) results for the chrominance components

Figure 3 Decoding results for sequence with high amount of intra coded areas with adaptive transform and non-adaptive spatial transform.



a) fixed transform

b) adaptive transform

Figure 4 Visual comparison of sequence where high-pass filtering of predicted signal at highest scale uses fixed or adaptive transform; decoded frame 93 of the "Soccer" sequence.