SHORT-TERM MOTION-BASED OBJECT SEGMENTATION

Marina Georgia Arvanitidou, Michael Tok, Andreas Krutz and Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Berlin Germany
{arvanitidou, tok, krutz, sikora}@nue.tu-berlin.de

ABSTRACT

Motion-based segmentation approaches employ either long-term motion information, which is computationally demanding, or suffer from lack of accuracy when employing short-term information. We present an automatic motion-based object segmentation algorithm for video sequences with moving camera, employing short-term motion information solely. For every frame, two error frames are generated using motion compensation. They are combined and a thresholding segmentation algorithm is applied. Recent advances in the field of global motion estimation enable outlier elimination in the background area, and thus a more precise definition of the foreground is achieved. We propose a simple and effective error frame generation and consider spatial error localization. Thus, we achieve improved performance compared with a previously proposed short-term motion-based method and provide subjective as well as objective evaluation.

Index Terms— object segmentation, camera motion estimation, thresholding

1. INTRODUCTION

Object segmentation in video sequences is a necessary and critical step for many applications such as video surveillance, object-based video coding and interactive multimedia services. Motion is among the salient characteristics that the human visual system perceives. Thus, it comprises a very powerful feature that the image processing community has adopted in order to address object segmentation tasks. In the cases of video sequences with moving camera, motion compensation is a prerequisite for motion-based segmentation. Hence, the parametric transforms that describe global (i.e. mainly induced by camera) motion between two video frames have to be calculated.

In the literature, motion-based segmentation algorithms rely on long-term motion information [1], [2] or on short-term motion information [3], [4]. Since the motion information contained in two or three frames is not always sufficient enough to differentiate the foreground from the background object, long-term motion-based algorithms have been proposed to overcome this issue. In these cases, the motion is calculated over multiple (more than three) frames and finally the background image that summarizes the aligned frames (sprite) can be accurately differentiated from the foreground. Nevertheless, long-term motion-based segmentation algorithms are very demanding in time or computational complexity. They have to process a large number of frames for the background modeling and thus they cannot be employed when real-time processing is a prerequisite.

On the other hand, short-term motion-based algorithms [5], [6] use one or two adjacent frames, to perform global motion estimation, compensation and segmentation on the resulting error frames. They are computationally less demanding than long-term motion-based segmentation approaches, but less accurate. It is critical to recognize and discard outliers in the background, and in the same time maintain the foreground object uncorrupted. Mech et al. combine in [3] the segmentation masks of two adjacent frames in a way that addresses well the background noise problem, at the cost of losing foreground information.

Here, we propose a short-term segmentation algorithm that can deal well with background outliers and yet maintain the foreground. Recently, advances in the area of global motion estimation were presented in [7], where a robust algorithm based on the Helmholtz principle was introduced. We employ this algorithm for global motion estimation and generate compensated error frames. We propose to combine the generated error frames in a simple and efficient way and then apply an enhancement of the thresholding segmentation algorithm presented in [8].

The remainder of this paper is organized as follows; Section 2 describes the employed global motion estimation approach. Section 3 overviews the system, presents the error frame generation and the segmentation algorithm applied on them. Experimental results are shown in Section 4 and Section 5 concludes the paper with discussion on perspectives for further work.

This work is developed within the PetaMedia project, a European Network of Excellence project, funded by the seventh European Research Framework Programme, contract FP7-21644.

978-1-61284-350-6/11/$26.00 ©2011 IEEE

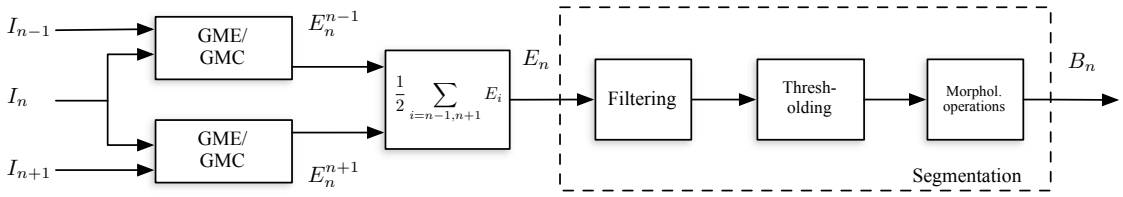


Fig. 1. Proposed system overview (GME/GMC = Global Motion Estimation/Compensation).

2. GLOBAL MOTION ESTIMATION USING THE HELMHOLTZ TRADEOFF ESTIMATOR

The employed global motion estimation algorithm is proposed in [7] and overviewed in Figure 2. The algorithm is based on the motion vectors that are calculated for every pair of adjacent frames of the video sequence. Authors in [7] show that this block based approach outperforms current pixel-based state of the art global motion estimation algorithms, maintaining very low computational complexity. The key issue is the removal of motion vector outliers using a robust estimator based on the Helmholtz principle, which can estimate parametric models from motion vector sets that have up to 80% of outliers.

In the first step a 4-parameter model (equation 1) that describes translation, rotation and zoom is applied, in order to reject the majority of outliers.

$$H = \begin{pmatrix} m_0 & m_1 & m_2 \\ -m_1 & m_0 & m_3 \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

Next, the remaining inliers are used to calculate an 8-parameter (perspective) motion model

$$H = \begin{pmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ m_6 & m_7 & 1 \end{pmatrix} \quad (2)$$

The perspective model is sufficiently precise to represent motion transformation in most natural video sequences. The calculation of a more simple parameter model, before calculating the more complicated perspective model, results in significant reduction of the computational burden without lack in accuracy.

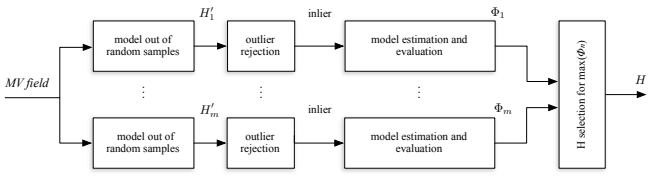


Fig. 2. Global motion estimation algorithm using the Helmholtz Tradeoff Estimator and two motion models

The procedure is performed for every subset m of motion vectors and all inliers are used to calculate a final model H_s by least squares for every subset. At the last stage, the final model is selected to be the one corresponding to the subset with the highest amount of inliers and lowest variance, according to $\Phi_s = I_s/\sigma_s'$ where I_s is the amount of inliers and σ_s' their estimation error standard deviation. In this way, the model can be estimated reliably even in an environment with a high amount of outliers.

3. SEGMENTATION OF ERROR IMAGES

The overview of the proposed system is shown in Figure 1. For the n^{th} frame of the video sequence we employ two adjacent frames, I_{n-1} and I_{n+1}, one for each direction in time. Two error frames E_n^{n-1} and E_n^{n+1} respectively are calculated after global motion estimation (GME) and global motion compensation (GMC). The error frames are combined in E_n and the thresholding segmentation step is applied on it. The system results finally in a binary image B_n where every pixel is labeled as either foreground or background. Following the error frame generation and segmentation steps are detailed.

3.1. Error frame generation

After global motion is estimated, global motion compensation follows. Ideally, the motion of the background (i.e. motion of the camera) is estimated exactly and the compensated frame differs from the original one only in regions with foreground motion. In practice, the background estimation is not perfect and it results in misclassifications of background pixels (false negatives). This can be addressed employing a thresholding algorithm.

For the n^{th} frame of the video sequence we calculate two error frames E_n^{n-1} and E_n^{n+1}. The two error frames are combined as

$$E_n = \frac{1}{2} \sum_{i=n-1,n+1} E_n^i \quad (3)$$

		Proposed			Reference		
Sequence	**Thresholding**	**P**	**R**	**F**	**P**	**R**	**F**
Stefan	Otsu [9]	0.68	0.73	**0.67**	0.84	0.44	0.50
	weighted mean [8]	0.72	0.70	**0.70**	0.89	0.44	0.56
	hysteresis	0.69	0.82	**0.73**	0.90	0.53	0.63
Race	Otsu [9]	0.70	0.79	**0.74**	0.88	0.44	0.56
	weighted mean [8]	0.70	0.78	**0.73**	0.87	0.49	0.62
	hysteresis	0.63	0.89	**0.74**	0.86	0.59	0.68
Biathlon	Otsu [9]	0.81	0.79	**0.80**	0.85	0.48	0.60
	weighted mean [8]	0.78	0.83	**0.80**	0.84	0.61	0.70
	hysteresis	0.75	0.85	**0.80**	0.82	0.69	0.74
Mountain	Otsu [9]	0.83	0.83	**0.83**	0.95	0.61	0.74
	weighted mean [8]	0.82	0.84	**0.83**	0.95	0.61	0.74
	hysteresis	0.76	0.91	**0.83**	0.92	0.73	0.81
Allstars	Otsu [9]	0.82	0.52	**0.62**	0.85	0.38	0.51
	weighted mean [8]	0.80	0.56	**0.65**	0.84	0.40	0.53
	hysteresis	0.73	0.66	**0.68**	0.80	0.48	0.59

Table 1. Mean precision (P), recall (R) and F-measure (F) for the reference and proposed algorithms for various thresholding methods.

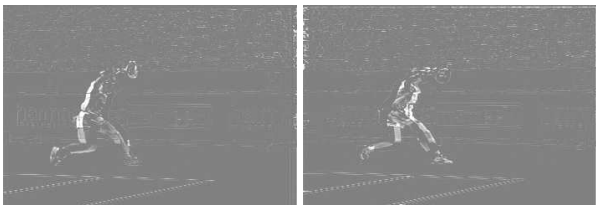
and the thresholding segmentation algorithm is then applied on E_n. In this way we get a more complete estimation of the foreground (high recall values), while having a very accurate background estimation we also obtain high precision values. The use of only one adjacent frame (e.g. I_{n-1}) would result in partial foreground detection, only at the edges of the motion direction. This is why Mech et. al. first proposed in [3] to use error frames from both directions. They applied a segmentation algorithm on the error frames E_n^{n-1} and E_n^{n+1} separately and then performed a logical AND operation between the resulting binary masks B_n^{n-1} and B_n^{n+1}. The AND operation ensured that foreground misclassifications are drastically reduced (high precision) at the the resulting B_n mask, but at the same time excluded significant amount of regions from the foreground object (low recall values). This shortcoming affected the overall segmentation quality in a bad manner as we show in Section 4. We overcome this issue by avoiding AND and including information from both error frames in the way described above.

3.2. Segmentation

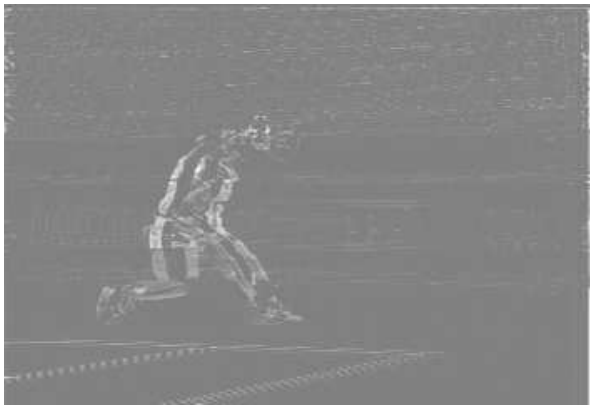
The error frames are intensity grayscale images. In the segmentation literature, thresholding algorithms are among the most efficient algorithms that address the intensity image segmentation problem. Krutz et al. presented in [8] an algorithm that employs in succession a) anisotropic diffusion filtering, b) weighted mean thresholding and c) morphological operations on the binary frames. An enhanced version of this approach is employed in the work presented here. Anisotropic diffusion filtering is used for reduction of high frequencies noise due to misestimations in the background. The weighted mean thresholding is given by

$$T(w) = w \cdot max(E'_n) + (1 - w) \cdot mean(E'_n) \qquad (4)$$

where E'_n is the normalized filtered error frame and w is a tuning constant. Finally, opening and closing morphological

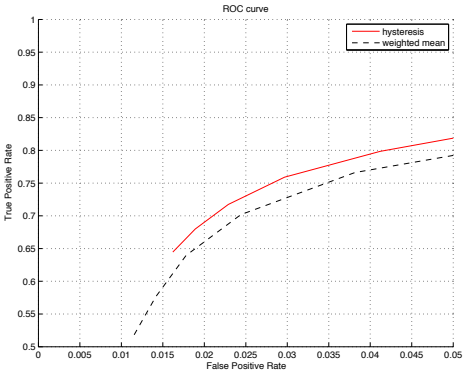


(a) E_{196}^{197} (b) E_{196}^{195}

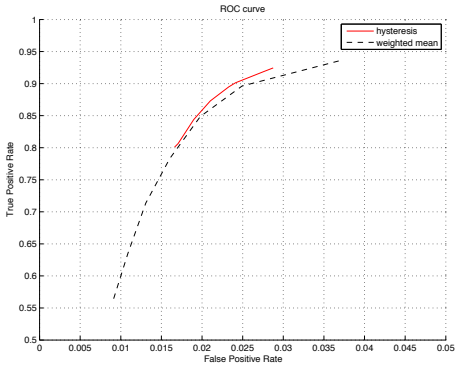


(c) E_{196}

Fig. 3. *Stefan* sequence, example error frames.



(a) *Stefan* sequence



(b) *Race* sequence

Fig. 4. Receiver operating characteristic curves

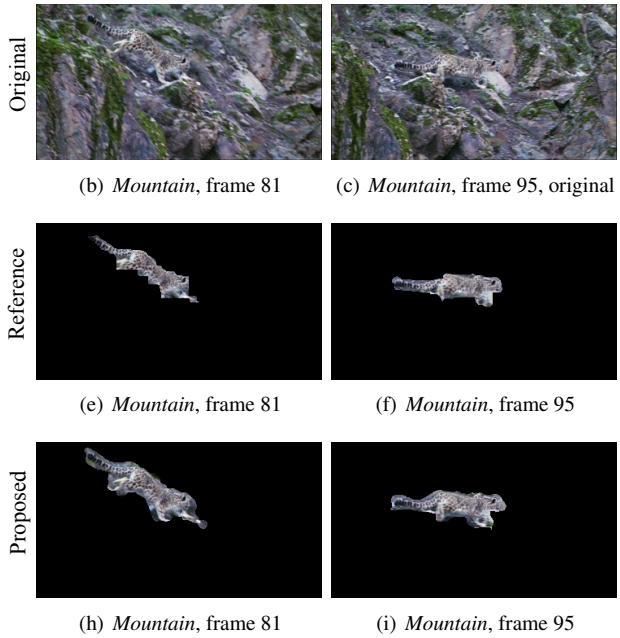
operations are employed for refinement of the binary segmentation mask by rejecting outliers.

The weighted mean thresholding in equation 4 adapts the threshold for every frame, but does not take into account the error localization. In Figure 3 we show error frame examples of one of our test sequences. There are significant error values in the foreground area and errors resulting from misestimations in the background area. At first stage, pixels with high error values should be labeled as foreground (F_0 region). Following, pixels with lower error values, that are spatially connected with F_0, should be favored against the ones not connected with F_0, even when the latter have high error values. Thus we employ two thresholds (*hysteresis* [10]). We begin by applying a low threshold $T(w_{low})$ using equation 4 for w_{low}. This results in a high amount of false positives, but we are fairly sure that most regions of the foreground are correctly set. We then apply a higher threshold $T(w_{high})$ that will be applied only on regions that are connected with the binary result from $T(w_{low})$. Once this process is complete we have a binary image where each pixel is marked as either foreground or background. The selection of the tuning constant will be discussed in the next section.

4. EXPERIMENTAL EVALUATION

Five test sequences are considered for experimental evaluation. The sequences are *Stefan* (sif, 300 frames), *Biathlon* (cif, 200 frames), *Race* (544 x 336, 100 frames), *Mountain* (352 x 192, 100 frames) and *Allstars (cif, 250 frames)*. To our knowledge, there is a lack in objective evaluation of short-term motion-based object segmentation algorithms in the literature. Therefore, in order to evaluate the performance of the proposed algorithm we present subjective as well as objective evaluation results, using manually created ground-truth sequences.

We compare the proposed algorithm (Proposed) with the algorithm inspired by [3] (Reference) which is described in Section 3.1. Table 1 shows precision (P), recall (R) and F-measure (F) values. Precision indicates how exact the segmentation is, meaning how accurately the background is estimated whereas recall shows how complete the foreground segmentation is. F-measure is the harmonic mean of precision and recall and is widely used as an objective overall indication of the segmentation quality. The global motion estimation algorithm described in Section 2 is used for both algorithms in order to have a fair comparison. Regarding thresholding, we evaluated each algorithm using the well known Otsu method [9], the weighted mean [8] and the enhancement of it, using hysteresis. Our proposed algorithm outperforms the reference one in any case for a given thresholding method. The improvement achieved using hysteresis thresholding is up to

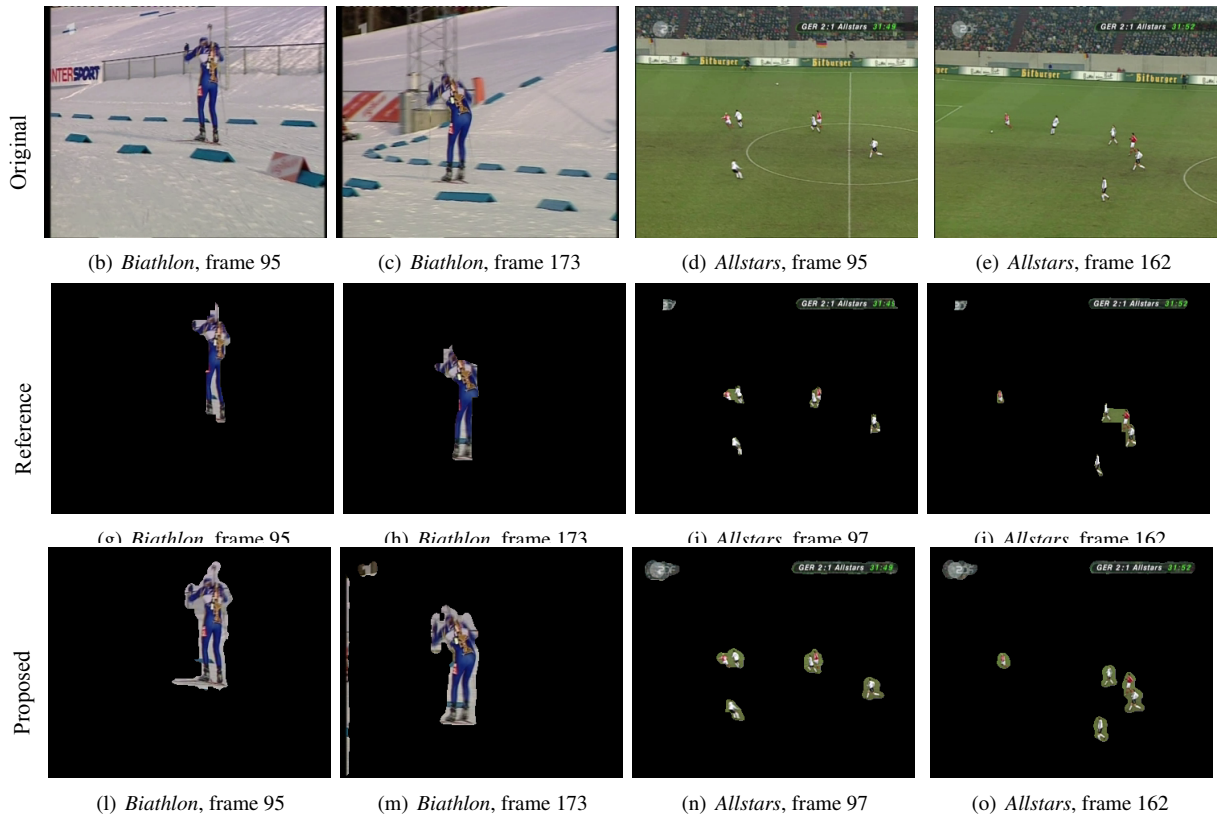


(b) *Mountain*, frame 81 (c) *Mountain*, frame 95, original

(e) *Mountain*, frame 81 (f) *Mountain*, frame 95

(h) *Mountain*, frame 81 (i) *Mountain*, frame 95

Fig. 5. Segmentation examples for the *Mountain* sequence, using the reference and the proposed algorithm (second and third row respectively).



(b) *Biathlon*, frame 95 (c) *Biathlon*, frame 173 (d) *Allstars*, frame 95 (e) *Allstars*, frame 162

(g) *Biathlon*, frame 95 (h) *Biathlon*, frame 173 (i) *Allstars*, frame 97 (j) *Allstars*, frame 162

(l) *Biathlon*, frame 95 (m) *Biathlon*, frame 173 (n) *Allstars*, frame 97 (o) *Allstars*, frame 162

Fig. 6. Segmentation examples for *Biathlon* and *Allstars* sequences, using the reference and the proposed algorithm (second and third row respectively).

10% for the case of *Stefan* sequence.

In order to investigate the performance of the constant w, used in the binarization step (Section 3.2), Receiver Operating Characteristic curves (ROC curves) are employed. ROC curves present the true positive rate against the false positive rate as the tuning constant w alternates. As described in [11] one point in ROC space is better than another if it is to the northwest of the first. In Figure 4 the continuous line corresponds to hysteresis thresholding for various (w_{high}, w_{low}) combinations, while the dashed line corresponds to weighted mean thresholding for various w values. Through this procedure we selected the constants w, w_{high} and w_{low} for best F-measure case and 0.1, 0.25 and 0.05 respectively are a satisfactory selection for all the test sequences.

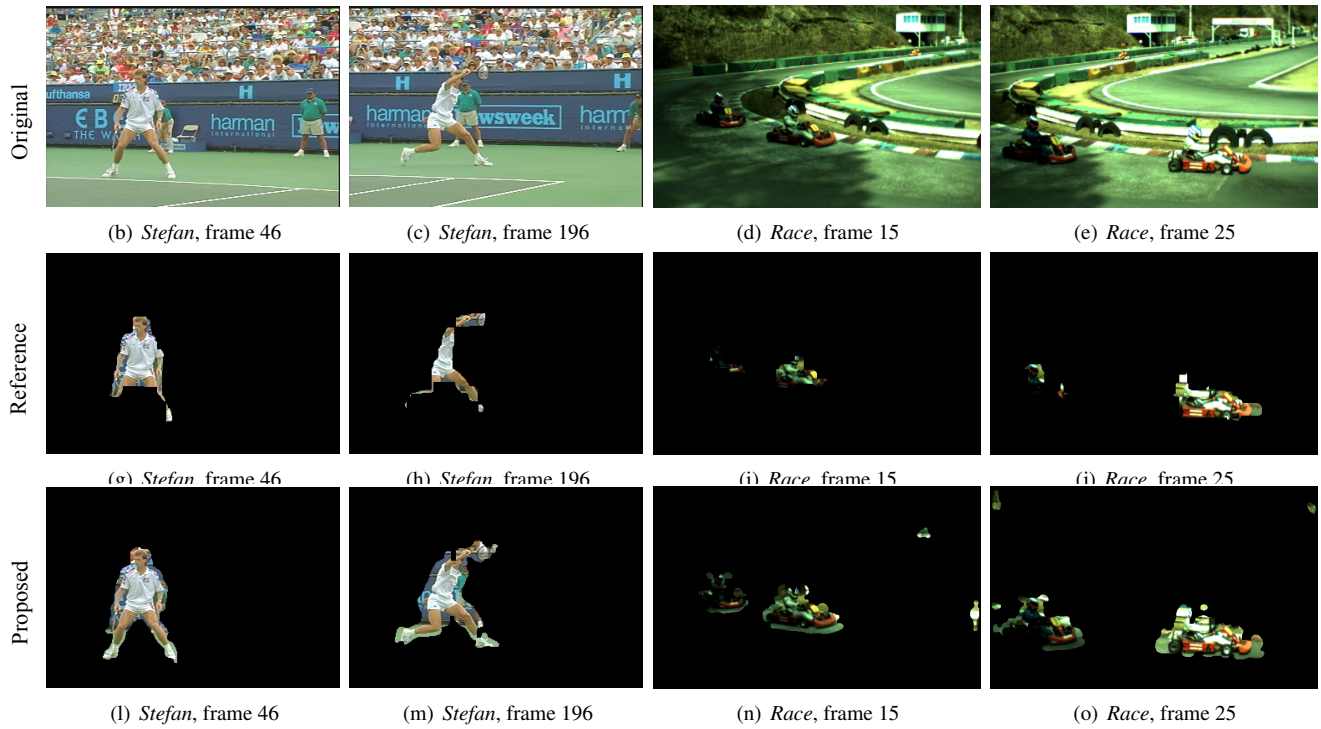
In Figures 5 - 7 we present example frames of the tested sequences using the reference and the proposed algorithm (second and third rows respectively). As it can be seen, the proposed algorithm results in more complete foreground detection However, in some cases, e.g. as depicted in Figure 7(o), the proposed algorithm results in increased amount of false positives. This is mainly due to large foreground object displacement from frame to frame, and does not affect significantly the overall results.

5. SUMMARY AND FURTHER WORK

An automatic motion-based object segmentation algorithm for video sequences with moving camera is presented. For every frame of the video sequence, only its predecessor and successor frames are employed. After motion estimation and compensation, the corresponding error frames are combined and the resulting error frame is subsequent to denoising filtering, thresholding and morphological operations. Objective experimental evaluation on five sequences with arbitrarily moving objects shows that we can outperform a previously proposed method. Further work includes consideration of color as well as spatiotemporal information, for increasing robustness and accuracy.

6. REFERENCES

[1] A. Krutz, A. Glantz, T. Borgmann, M. Frater, and T. Sikora, "Motion-based object segmentation using local background sprites," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 1221–1224.



(b) *Stefan*, frame 46 (c) *Stefan*, frame 196 (d) *Race*, frame 15 (e) *Race*, frame 25

(g) *Stefan*, frame 46 (h) *Stefan*, frame 196 (i) *Race*, frame 15 (j) *Race*, frame 25

(l) *Stefan*, frame 46 (m) *Stefan*, frame 196 (n) *Race*, frame 15 (o) *Race*, frame 25

Fig. 7. Segmentation examples for *Stefan* and *Race* sequences, using the reference and the proposed algorithm (second and third row respectively).

[2] M. Bhaskaranand and S. Bhagavathy, "Motion-based object segmentation using frame alignment and consensus filtering," in *Proceedings of International Conference on Image Processing*, 2010.

[3] R. Mech and M. Wollborn, "A noise robust method for segmentation of moving objects in video sequences," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 1997, vol. 4, pp. 2657–2660.

[4] Mao Ling and Xie Mei, "Automatic segmentation of moving objects in video sequences based on spatio-temporal information," in *Proceedings of the International Conference on Communications, Circuits and Systems*, Jul. 2007, pp. 750–754.

[5] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: a region labeling approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 7, pp. 597–612, Jul. 2002.

[6] M.G. Arvanitidou, A. Glantz, A. Krutz, T. Sikora, M. Mrak, and A. Kondoz, "Global motion estimation using variable block sizes and its application to object segmentation," in *Proceedings of the international Workshop on Image Analysis for Multimedia Interactive Services*, May 2009, pp. 173–176.

[7] M. Tok, A. Glantz, M. G. Arvanitidou, A. Krutz, and T. Sikora, "Compressed domain global motion estimation using the helmholtz tradeoff estimator.," in *Proceedings of the IEEE International Conference on Image Processing*, Hong Kong, Sep. 2010.

[8] A. Krutz, M. Kunter, M.l Mandal, M. Frater, and T. Sikora, "Motion-based object segmentation using sprites and anisotropic diffusion," in *Proceedings of the international Workshop on Image Analysis for Multimedia Interactive Services*, Jun. 2007.

[9] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62 –66, 1979.

[10] Paul Rosin and Tim Ellis, "Image difference threshold strategies and shadow detection," in *Proceedings of the British Machine Vision Conference*, 1995, pp. 347 – 356.

[11] Tom Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters, Elsevier*, vol. 27, pp. 861–874, Jun. 2006.