Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 792028, 19 pages
doi:10.1155/2008/792028

Research Article

Joint Wavelet Video Denoising and Motion Activity Detection in Multimodal Human Activity Analysis: Application to Video-Assisted Bioacoustic/Psychophysiological Monitoring

C. A. Dimoulas, K. A. Avdelidis, G. M. Kalliris, and G. V. Papanikolaou

Laboratory of Electroacoustics and TV Systems, Department of Electrical and Computer Engineering, Laboratory of Electronic Media, Department of Journalism and Mass Communication, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Correspondence should be addressed to C. A. Dimoulas, babis@eng.auth.gr

Received 28 February 2007; Revised 31 July 2007; Accepted 8 October 2007

Recommended by Eric Pauwels

The current work focuses on the design and implementation of an indoor surveillance application for long-term automated analysis of human activity, in a video-assisted biomedical monitoring system. Video processing is necessary to overcome noise-related problems, caused by suboptimal video capturing conditions, due to poor lighting or even complete darkness during overnight recordings. Modified wavelet-domain spatiotemporal Wiener filtering and motion-detection algorithms are employed to facilitate video enhancement, motion-activity-based indexing and summarization. Structural aspects for validation of the motion detection results are also used. The proposed system has been already deployed in monitoring of long-term abdominal sounds, for surveillance automation, motion-artefacts detection and connection with other psychophysiological parameters. However, it can be used to any video-assisted biomedical monitoring or other surveillance application with similar demands.

1. INTRODUCTION

Video surveillance is a common task in human biomedical monitoring applications, especially for prolonged recording periods, where physical supervision is not feasible [1]. Its utilization usually involves (a) surveillance of human behavior/anxiety in combination with various other psychophysiological parameters, (b) continuous monitoring in critical health-care environments or in cases of subjects that need special treatment for safety reasons (neonatal, handicaps, elderly people, etc.), (c) detection and isolation of movement artefacts that affect the integrity of the psychophysiological data, (d) validation and verification of various health-related symptoms/events, such as cough, apnoea episodes, restless leg syndrome, and so forth [1–7]. The majority of the video-assisted biomedical monitoring systems are engaged in polysomnography recordings during sleep studies [2–7], in various neurophysiology and kinesiology-related studies [8–10], for the extraction of temporal motion strength signals from video recordings of neonatal seizures [11]. Video monitoring and analysis allows physicians to evaluate the exact experimental condition under which the biomedical data were acquired [1]. The method described in this paper was employed in long-term gastrointestinal motility monitoring by means of abdominal sounds [1, 12], to offer an alternative approach in detecting and rejecting motion-produced sliding noises; it was also very helpful during evaluation of audio-based automated pattern recognition, which offered an alternative approach in artefacts detection and removal [1, 13]. Besides these two technical aspects, the incorporation of video surveillance was decided in order to be able to correlate the phases of the gastrointestinal bio-acoustic activity with other physiological parameters previously mentioned, such as brain-activity, sleep cycles' alteration, respiratory-related parameters, or even abnormal behavior caused by psychological factors [1].

Most of the video-assisted biomedical applications are dealing with the fact that nonoptimal capturing conditions are unavoidable, since lighting the scene in the adequate illumination-levels would produce discomfort to subjects,

affecting the validity of the experimental psychophysiological monitoring procedure [1–7]. In addition, overnight recordings are conducted in sleep laboratories or in other biomedical examinations, including our gastrointestinal motility monitoring application [1, 12]. As a result, low-light cameras, night vision, and infrared devices are engaged in most cases, worsening the noise contamination problems that are usually met in general video monitoring applications. Therefore, video denoising processing is necessary for enhancement of the captured image-sequences to improve perceptual analysis during the examination of the content.

Apart from video enhancement, motion detection and synchronization of the surveillance data with the acquired psychophysiological parameters are quite common in most video-assisted biomedical applications [1, 4, 8–11]. Except from the enhancement aspects, noise removal is essential for all the involved video processing stages, such as compression, motion detection/estimation, object segmentation/characterization, and so forth [1, 14–18]. Another important issue that needs careful treatment, especially for prolonged surveillance periods, is the ability to automate indexing, characterization, and summarization of the captured audio-visual content, facilitating easy browsing, searching, and retrieval [1, 19–24]. Video motion detection is one of the most applicable techniques usually employed to track changes in the monitored area, offering also the ability to extract summarization plots and pictures [1, 24–29]. This is the reason that the MPEG-7 protocol incorporates various motion descriptors for content management purposes [19–21].

Summing up, the purpose of the current work is to provide an integrated solution for video enhancement, event detection, and summarization of long-term surveillance content, which has been acquired under suboptimal capturing conditions. Spatiotemporal wavelet Wiener filtering denoising techniques are considered in combination with wavelet-adapted motion detection algorithms, to deal with the demands of video enhancement and efficient content indexing/description. These demands are quite common to most video surveillance systems, regardless the type of their utilization, for example, biomedical monitoring, security systems, traffic monitoring, human machine interaction, and so forth. Thus, the proposed methodology can be applied to any of these areas.

The paper is organized as follows. The problem definition is described in Section 2. State of research and related methods are presented in Section 3, providing a quick overview of contemporary video denoising approaches, motion detection techniques, and recent strategies in audio-visual content description/management. The proposed methodology is analyzed in Section 4. Experimental results are discussed in Section 5, where evaluation of the proposed methods is carried out in combination with conclusion and future work remarks.

2. PROBLEM DEFINITION

Noise contamination is a typical problem to most electronic communication systems, including surveillance applications. In most of the cases, video enhancement by means of noise reduction is necessary in order to improve image quality, increase compression efficiency, and facilitate all video processing stages that may possibly follow [14–18]. For example, by applying simple order-statistics filters in effort to reduce noise, an improvement in compression efficiency by a factor 1.5 to 2 was observed, without the presence of noticeable compression artefacts [1]. This is explained by the fact that the presence of noise might be interpreted as excessive and random motion, deteriorating the compression efficiency of the related motion-compensation algorithms [14–18, 27]. In addition, erroneous motion estimation (ME), usually expressed by motion vectors (MVs), may occur [14, 27]. This has a negative impact on background/foreground segmentation (BRFR) results, usually involved in surveillance systems [1, 25, 26, 28].

Video signals can be corrupted by noise during acquisition, recording, digitization, processing, and transmission. Typical examples of video-noise include CCD-camera noise, analog channels interferences, magnetic-recording noise, quantization noise during digitization, and so forth [14–18]. According to [15], in digital cameras the video noise level may increase because of the higher sensitivity of the new CCD cameras and the longer exposures. In general, the noise signal can be modelled as stochastic process, which is additive or multiplicative, signal-dependent or independent, white or colored, according to its spectral properties [15]. Most researchers tend to model the above types of video-noise sources as independent identically distributed additive and stationary zero-mean noise, which is the simplest Gaussian additive white noise model described from the following equation [14–18]:

$$I_X(i, j, n) = I_S(i, j, n) + I_N(i, j, n), \tag{1}$$

where I_X is the luminance of the noise contaminated image, I_S the noise-free image, I_N the 2D noise signal, i, j are the spatial indexes, and n the time-index for the images sequences (frame number). Equation (1) suggests that only grey-scale images are considered, since I_X, I_S, I_N refer to the intensities of the corresponding colorless 2D signals. This model was also adopted in the current work, mainly due to the fact that colored video increases the computational load, without increase of the usefulness of the provided information. Additionally, night vision equipment inherently belongs to monochromatic video systems, so that greyscale images were selected to allow similar treatment in both diurnal and nocturnal surveillance. However, (1) can be extended to the appropriate color space components to apply on color video cases. To answer the noise contamination problem, most video denoising algorithms tend to employ 2D image (spatial) filtering, motion detection, and temporal smoothing.

A consequent problem is the erroneous estimation of the background image $B(i, j, n)$. The noised versions of both the intensity and the background images deteriorate the efficiency in the estimation of the foreground objects, usually extracted via the subtraction of the previously mentioned signals $I_X(i, j, n)$ and $B(i, j, n)$. To deal with the stated problem, there is a necessity for algorithms that can effectively accomplish the BRFR segmentation task under the presence of nonoptimal conditions, previously discussed. Among

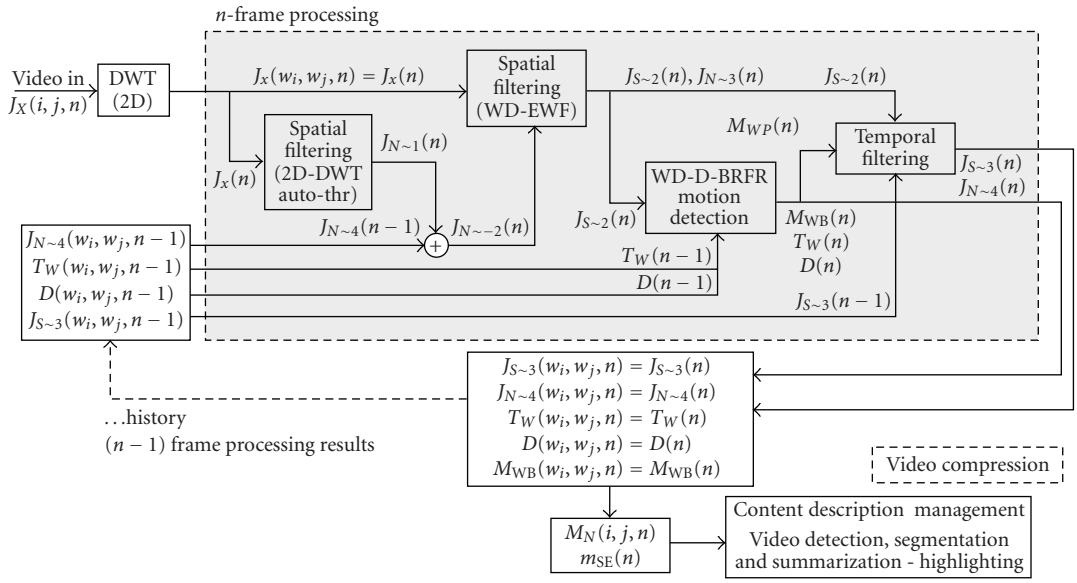


Figure 1: Block diagram of the JWVD-MAD algorithm.

the wanted characteristics of those algorithms is the ability to accurately extract suitable motion parameters that could be consequently used for content management purposes [1, 25–28], especially for prolonged monitoring periods. Thus, motion-detection-based video indexing is quite useful in surveillance applications, while the interaction with audio content and other modalities can serve as a powerful tool towards multimodal event detection segmentation and summarization [1, 12, 13].

3. RELATED RESEARCH AND THE SELECTED APPROACHES

A quick overview of the research background in video denoising, video-motion detection, and audio-visual content management is needed before the proposed techniques are further analyzed. This paragraph mainly focuses on the methods that are utilised in the current work.

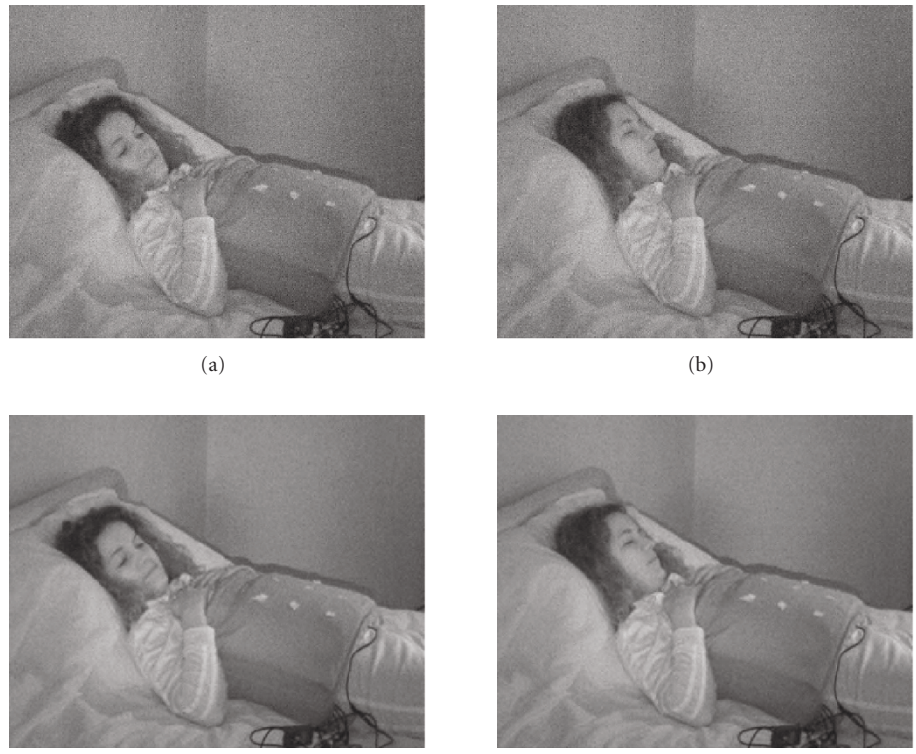
3.1. *Video denoising overview*

Based on the remarks of the previous paragraph, most video denoising/enhancement algorithms implement temporal, spatial, and spatiotemporal filtering, to take advantage of the corresponding redundancy (similarities), usually met in natural video sequences [14–18]. The estimation of the noise variance $\sigma_N^2(n)$ is necessary in order to deploy spatial filtering techniques for noise suppression. Structural characteristics of the image morphology are also considered to avoid creating blurring at image edges [15, 16, 18]. Temporal smoothing, on the other hand, tends to produce motion-artefacts (blurring), when it is applied to moving regions. To face these difficulties, temporal smoothing is usually applied along with the estimated pixel-motion-trajectories [14, 18, 28].

As already stated in Section 2, the noise contamination problem is unavoidable in most electronic communication systems, including video applications. The unwanted effects of the video-noise presence have been already discussed and analyzed in most video denoising references [14–18]. Focusing on the demands of the current human-activity video-surveillance system, noise worsens the quality of the acquired images, produces erroneous estimations of the motion-activity parameters, and deteriorates the video compression efficiency. Video denoising, as it happens with all single-sided signal restoration techniques [14, 30, 31], try to estimate the noise statistical attributes from the available noise-contaminated signal, in order to apply spatiotemporal filtering. In addition, autonoise estimation methods have been proposed to facilitate unsupervised image and video denoising [14–18, 31–35]. Wiener filter, which minimizes the mean-square error between the original clean signal and the estimated one obtained during the reconstruction procedure, is the basis for the current denoising approach. Thus, extending the 1D processing case [30], the Wiener filtering operation in the frequency-space domain is described by the following equation [14, 31, 35]:

$$F_{S\sim}(\omega_i,\omega_j) = \begin{cases} \left[1 - c_{\mathrm{WF}} \cdot \dfrac{P_{N\sim}(\omega_i,\omega_j)}{P_X(\omega_i,\omega_j)}\right] \cdot F_X(\omega_i,\omega_j), \\ \qquad \text{if } c_{\mathrm{WF}} \cdot \dfrac{P_{N\sim}(\omega_i,\omega_j)}{P_X(\omega_i,\omega_j)} \le 1, \\ 0, \quad \text{otherwise}, \end{cases} \tag{2}$$

where $F_X(\omega_i,\omega_j)/F_S(\omega_i,\omega_j)/F_N(\omega_i,\omega_j)$ are the Fourier transforms of the noised $I_X(i,j)$/clean $I_S(i,j)$/noise $I_N(i,j)$



(a) (b) (c) (d)

Figure 2: Qualitative analysis of denoising results: (a)-(b) noised frames, (c)-(d) reconstructed frames.

images, and $P_X(\omega_i, \omega_j)/P_S(\omega_i, \omega_j)/P_N(\omega_i, \omega_j)$ are the corresponding power spectrum estimates. Equation (2) describes the so-called 2D parametric Wiener filter, where the c_{WF} parameter is used to control the amount of noise suppression and it may be omitted in the simplest case of classical Wiener filter ($c_{\mathrm{WF}} = 1$) [30, 31]. The "$\sim$" symbol, which is used in the $F_{S\sim}(\omega_i, \omega_j)$, $P_{N\sim}(\omega_i, \omega_j)$ components of (2) denotes that the corresponding signals are estimations of the original ones (clean image spectrum $\mathbf{F}_S$ and noise power $\mathbf{P}_N$), since the latter are not available. It is obvious that the estimated noise-free image $I_{S\sim}(i, j)$ can be obtained via inverse Fourier transform of the processed spectrum $F_{S\sim}(\omega_i, \omega_j)$.

Besides Fourier components, any other spectral analysis tool can be used in (2), including filter banks, subband decomposition, and wavelets. In the last case, the $F_X(\omega_i, \omega_j)/F_S(\omega_i, \omega_j)/F_N(\omega_i, \omega_j)$ components of (1) are replaced with the wavelet coefficients $J_X{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})/J_S{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})/J_N{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})$, where l denotes the decomposition level ($l = 1, 2, \ldots L_W$) and AD is the approximation/details index: AD = "*Low-Low*", "*Low-High*", "*High-Low*", "*High-High*" = {LL, LH, HL, HH}. The new power estimates $P_X{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})/P_S{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})/P_N{}^{(l;\mathrm{AD})}(w_{li}, w_{lj})$ are now referred to the "wavelet images" usually obtained via 2D discrete wavelet transform (DWT) and 2D wavelet packets (following the "subsampling by 2" rule at every wavelet decomposition node l), or even undecimated wavelet transform (UWT) [16–18, 32]. Wavelet shrinkage is deployed according to (3), while the noise-free image is estimated by applying inverse wavelet transform (IWT) to the processed coefficients:

$$J_{S\sim}(w_i, w_j) = \begin{cases} \left[1 - c_{\mathrm{WF}} \cdot \dfrac{P_{N\sim}(w_i, w_j)}{P_X(w_i, w_j)}\right] \cdot J_X(w_i, w_j), \\ \qquad \text{if } c_{\mathrm{WF}} \cdot \dfrac{P_{N\sim}(w_i, w_j)}{P_X(w_i, w_j)} \le 1 \\ 0, \quad \text{otherwise,} \end{cases} \qquad \forall (l; \mathrm{AD}) \tag{3}$$

omitting the corresponding indicators $(l; \mathrm{AD})$ for the sake of simplicity. This is to be followed throughout the rest of the paper for all the wavelet-based quantities, unless otherwise stated.

The above image processing equations may be also used for video Wiener denoising. As stated, the simplest approach to video denoising is to employ image filtering to every frame n of the video sequences. Thus, (2) and (3) may be used for the case of video spatial filtering, by replacing arguments (ω_i, ω_j) and (w_i, w_j) with (ω_i, ω_j, n) and (w_i, w_j, n), for each $(l; \mathrm{AD})$, respectively. This approach, however, does not take into consideration similarities between successive frames (temporal smoothing). On the other hand, we may consider that all the frequency/wavelet image components (pixels) of (2) and (3) are 1D curves versus time, so that 1D Wiener filtering could be applied to every single one of them (temporal-only smoothing: n is the only independent variable in the arguments of the previous equations) [14, 31].

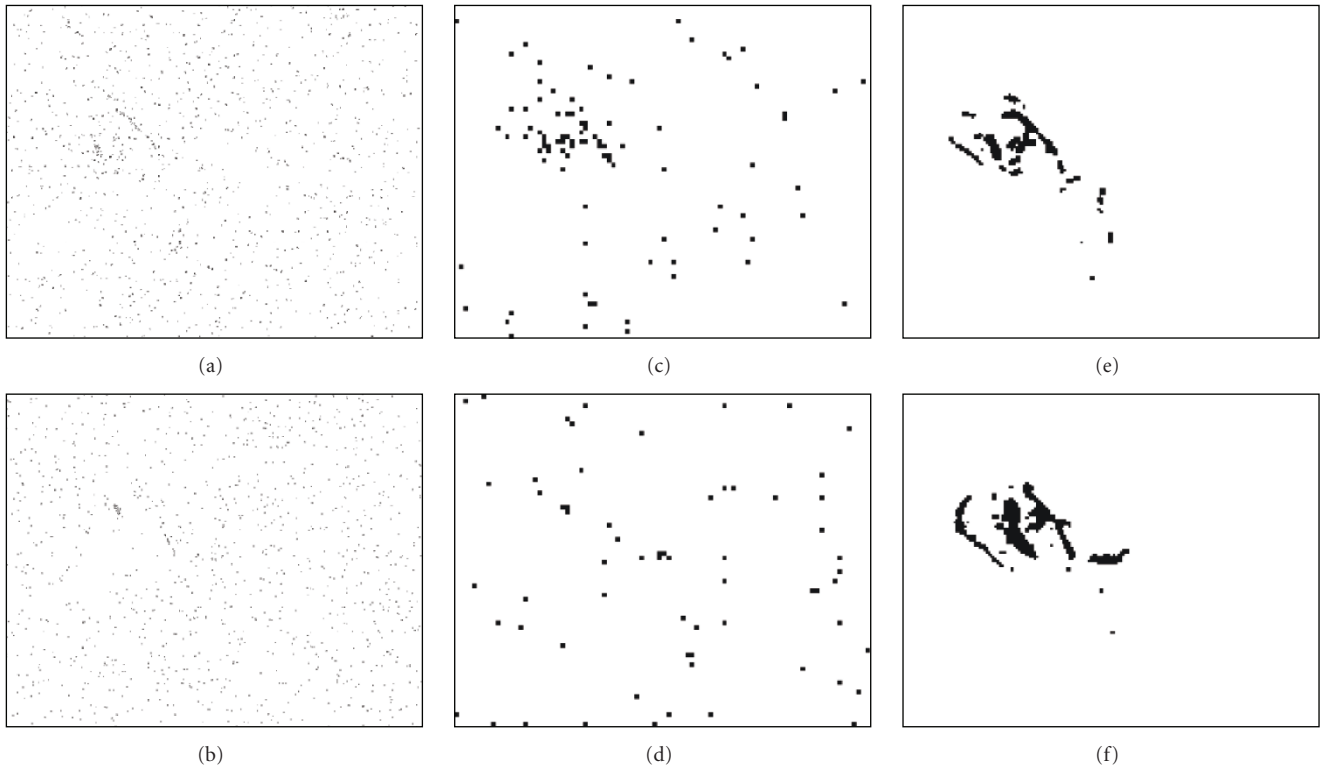


Figure 3: Qualitative analysis of motion detection results: (a)-(b) motion images extracted with the TD-BRFR method, (c)-(d) motion images extracted with the WD-BRFR method, (e)-(f) motion images extracted with the JWVD-MAD algorithm.

The appearance of motion artefacts in the case of moving pixels is a common disadvantage of these techniques, already discussed. There have been researchers in past works that have evaluated the order of operations (spatial and temporal filtering) that provides optimal de-noising [14, 18], while various motion compensation strategies have been proposed to reduce motion artefacts during temporal smoothing [14, 16, 18, 35]. Taking these facts into account, 1D and 2D wavelet domain Wiener filtering algorithms can be effectively combined to provide improved video denoising solutions. The so-called empirical Wiener filter [36] is another related issue concerning a strategy that was also adopted in the current work.

3.2. Video motion detection overview

Video motion detection plays a very important role in surveillance systems. In contrast to motion estimation techniques that try to compute MVs in order to find all the motion attributes, motion detection algorithms try to classify image-pixels to moving and nonmoving ones, so that they are usually computationally faster and easier to implement [22, 27]. There is an interaction between motion detection and motion estimation methods. In motion-compensated compressed video, MVs may be utilized to offer motion detection results. On the other hand, motion detection can be deployed as a preprocessing stage to facilitate motion estimation and to improve compression efficiency, an approach that is closer to the strategy adopted in the current work. Thus, considering the case that no MVs are available, motion detection is usually implemented via time differencing comparisons, optical flow techniques and background subtraction methods [25, 26]. We will focus on the last subcategory presenting the BRFR segmentation methods developed by Collins et al. [25] and Töreyin et al. [26], since they were used as the basis for the modified joint wavelet video denoising and motion activity detection (JWVD-MAD) algorithm, proposed in the current paper.

Collins et al. [25] developed a time-domain BRFR classification method (TD-BRFR) using exponential moving average techniques (ExpMA):

$$B(i,j,n+1) = \begin{cases} a_m \cdot B(i,j,n) + (1-a_m)\cdot I(i,j,n), \\ \qquad\qquad \text{if the } (i,j) \text{ pixel is nonmoving}, \\ B(i,j,n), \quad \text{otherwise}, \end{cases} \tag{4}$$

where the i, j indexes determine the images' spatial coordinates, the $n, n+1$ indexes determine the video frame number, a_m is the "motion-constant" utilized in the ExpMA BRFR procedure, $B(i,j,n)$ is the estimated background image at frame n, and $I(i,j,n)$ is the image intensity (greyscale image) at frame n, which is considered to be noise free. In order to be able to execute operations inside (4), the motion-pixel

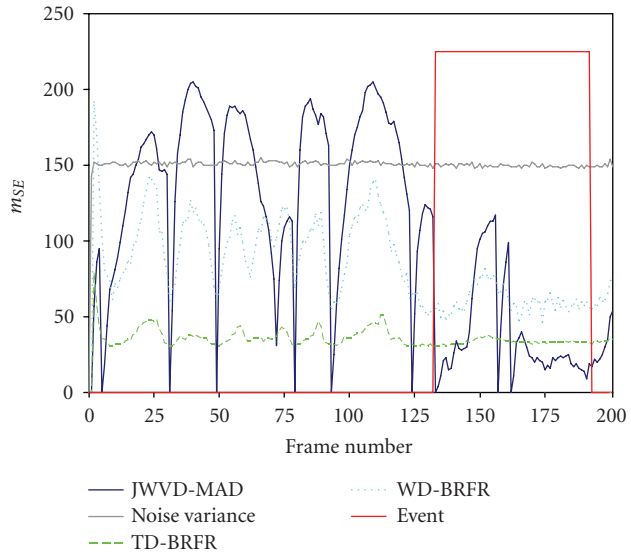


Figure 4: Motion activity curves for the example presented in Figure 3 using a threshold value equal to $T_{\text{event}} = 40$ (the estimated noise variance is plotted in grey color and the manual-tagged "head-turn" event is signed with red color; the slight event is detected as significant activity with the proposed methodology, in contrast to the baseline methods, where the motion curves m_{SE} are vanished at very low levels).

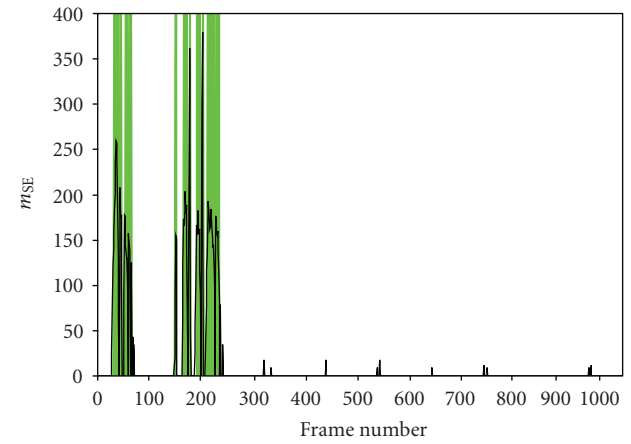


Figure 5: Motion activity curve and video motion detection results via the VDSS method ($T_{\text{event}} = 40$): the green-color curves represent the automatically detected events.

masks $M_P(i, j, n)$ are estimated at every frame n [1, 25, 26]:

$$M_P(i,j,n) = \left| I(i,j,n) - I(i,j,n-1) \right| > T(i,j,n). \quad (5)$$

The threshold parameter $T(i, j, n)$ is also adapted iteratively via the ExpMA procedure described in the following equation:

$$T(i,j,n+1) = \begin{cases} a_m \cdot T(i,j,n) + (1-a_m) \cdot c_m \cdot |I(i,j,n) - B(i,j,n)|, \\ \qquad \text{if the } (i,j) \text{ pixel is nonmoving}, \\ T(i,j,n), \quad \text{otherwise}, \end{cases} \quad (6)$$

where the "motion comparison" parameter c_m ($c_m > 1$) is used to control the motion detection sensitivity (the greater the c_m value, the lower the motion detection sensitivity). Equations (4), (5), and (6) are executed consequently, with the initial condition $B(i, j, 1) = I(i, j, 1)$. Additionally, the threshold parameter needs to be empirically defined at a constant value T_{const} during procedure initiation: $T(i, j, 1) = T_0$, for all i, j. The motion binary images $M_B(i, j, n)$ are finally computed as follows:

$$M_B(i,j,n) = \left| I(i,j,n) - B(i,j,n-1) \right| > T(i,j,n). \quad (7)$$

Töreyin et al. [26] proposed a wavelet domain BRFR segmentation (WD-BRFR), taking advantage of the available image wavelet coefficients $J(w_i, w_j, n)$. Thus, (4)–(7) may be employed in the wavelet domain by replacing image intensities $I(i, j, n)$ with the coefficients $J(w_i, w_j, n)$. Wavelet background images $D(w_i, w_j, n)$ are then estimated instead of $B(i, j, n)$, while subband binary motion images $M_{\text{WB}}(w_i, w_j, n)$ are calculated at the involved wavelet scales. A rescaling procedure is necessary to extract the final binary motion image $M_B(i, j, n)$, taking into account the subsampling grid employed during wavelet transform [26]. Specifically, the involved 2D motion coefficients $M_{\text{WB}}(w_i, w_j, n)$ are projected to the corresponding $M(i, j, n)$ motion matrices, and the final binary motion image M_B is generated via an OR Boolean function,

$$\begin{aligned} M(i,j,n) &= M\left([2^l w_i : 2^l w_i + 2^l - 1], [2^l w_j : 2^l w_j + 2^l - 1], n\right) \\ &= M_{\text{WB}}(w_i, w_j, n) \\ i &= [0, N_H - 1], \quad j = [0, N_V - 1], \\ w_i &= \left[0, \frac{N_H}{2^l} - 1\right], \quad w_j = \left[0, \frac{N_V}{2^l} - 1\right] \\ M_B(i,j,n) &= \text{OR}\{M(i,j,n)\}, \quad \forall (l; \text{AD}). \end{aligned} \quad (8)$$

Töreyin et al. [26] also suggested a second level for motion detection refinement, by lowering the thresholding criteria at pixels neighbouring to motion regions, taking structural aspects into account for object detection. Besides BRFR segmentation, no other wavelet processing was engaged, since both the images $I(i, j, n)$ and the corresponding wavelet coefficients $J(w_i, w_j, n)$ were considered to be noise free [26].

3.3. Audio-visual content management approaches

A common task in most audio-visual surveillance demanding applications is the implementation of effective content management tools in order to facilitate easy video browsing, indexing, searching, and retrieval. Within this context, various techniques have been developed for image similarity comparisons, video characterization, and abstraction via highlighting image sequences. In general we may distinguish two basic strategies: color information and motion-based parameters [19–21].

Color-based techniques tend to give better results, but they are more computationally demanding when compared to the motion-based approaches. Video motion techniques feature easier implementation and are preferred in surveillance applications, where color changes are difficult to follow [24, 25, 27]. Another advantage is that motion features can be implemented to colorless video and night vision image sequences.

Motion parameters are easily extracted from the MVs, available in MPEG streams or similar motion-compensated, compressed videos. A representative example is the MPEG-7 motion activity descriptor that uses statistical attributes of MVs (variance, spatial/temporal distribution) in order to describe the motion pace of video sequences. In the case that MVs are not available, motion estimation is usually employed via block matching algorithms. However, there are many cases (including surveillance applications) where motion detection is preferred (over motion estimation) and MVs are not applied, due to the easier implementation of the related algorithms. Thus, extending the analysis presented previously, binary motion images may be further utilized to extract 1D "motion-intensity curves" in order to facilitate video indexing and characterization [1, 22]. It is obvious that video sequences with intensive motion would result to a great number of moving points ($M_B(i, j, n) = 1$), while complete absence of moving pixels would be observed in the case of motionless video sequences.

4. THE PROPOSED JWVD-MAD METHODOLOGY

The proposed methodology aims to provide an integrated framework for surveillance video enhancement, event detection, and abstracting. Specifically, wavelet-domain motion detection is employed, as in the case of [26], using the iterative ExpMA scheme initially proposed in [25]. The main difference is that the current method is applied prior to final compression, considering the presence of additive contamination noise. In addition, we introduce the "active background" concept, since the still images, considered as background, are stabilized to new "backgrounds" once the detected movement is completed. Within this context, a dynamic BRFR segmentation procedure (WD-D-BRFR) is initialized each time a motion event is terminated. A block diagram describing all the processing phases of the proposed methodology is presented in Figure 1.

The BRFR segmentation algorithms presented in the previous paragraph [25, 26] did not take into account video degradation issues due to the presence of noise. Thus, $I(i, j, n)$ and $J(w_i, w_j, n)$ of (4)–(6) need to be replaced with the $I_S(i, j, n)$ and $J_S(w_i, w_j, n)$. However, these original noise-free signals are not available due to noise contamination problem and the noised versions $I_X(i, j, n)$ and $J_X(w_i, w_j, n)$ should be used instead. The current method proposes the use of the denoised signals $I_{S\sim}(i, j, n)$ and $J_{S\sim}(w_i, w_j, n)$, where, as already mentioned, the "$\sim$" symbol expresses the fact that the noise-free estimated signals are not identical to the original ones. This indexing approach is also used for the estimated noise signals in the space or the wavelet domain: $I_{N\sim}(i, j, n)$ and $J_{N\sim}(w_i, w_j, n)$, respectively.

4.1. Video denoising by means of spatiotemporal wavelet filtering (VD-STWF)

The first step in the proposed JWVD-MAD methodology is the deployment of wavelet filtering in order to obtain the noise-free estimations of the available signals. Since both temporal filtering and spatial filtering are engaged in succession, there are differences between the various noise/signal estimations denoted by "$\sim$". To deal with this "notation difficulty" we decided to define the number of filtering procedures employed for a specific estimation, next to the "$\sim$" symbol. For example, the $I_{N\sim1}(i, j, n)$ parameter indicates that the current noise estimation has been produced via a single denoising process (i.e., spatial filtering), while the $I_{N\sim2}(i, j, n)$ value is estimated after the insertion of a second denoising process (i.e., temporal smoothing). In any case, both temporal smoothing and spatial filtering are implemented directly in the wavelet domain, to take advantage of the wavelet-based video denoising advantages [16–18]. Thus, the WD-BRFR approach, initially proposed by Töreyin et al. [26] will be followed, allowing direct use of the processed wavelet coefficients $J_{S\sim}(w_i, w_j, n)$, without the necessity of applying IWT (if no other processing is involved). This is also beneficial in the case that a wavelet compression algorithm is followed.

Let us turn our attention to the block diagram of Figure 1. It is obvious that spatial filtering precedes temporal smoothing, with the last one to be implemented after motion detection for artefacts (blurring) avoidance. However, temporal similarities are also exploited during the estimation of the noise power coefficients $P_N(w_i, w_j, n)$. Considering that noise energy characteristics do not change very rapidly, noise history can be used for the refinement of the wavelet thresholding rules. Wavelet image denoising is additionally applied for noise estimation at the current frame (n). In general, any 2D wavelet autothresholding method can be employed to this preprocessing step of the empirical Wiener filter [36]. The soft-thresholding version using the parametric threshold of "$\mathrm{Th}_N = k_m \cdot \sigma_N$" was finally selected (by introducing the multiplicative factor k_m), since it proved to best combine efficiency with reduced complexity.

There are applications [36] where empirical Wiener filtering has been implemented in the wavelet domain for video denoising purposes. However, the approach followed in this paper is quite different from the method proposed in [36], where autothresholding results are used to estimate SNR in order to reconfigure Wiener filter for a second wavelet processing scheme. In the current work, we avoid to perform IWT by using the exact wavelet topology in both denoising stages (autowavelet shrinkage via soft thresholding and wavelet Wiener filtering). In addition, we introduce the wavelet noise power that has been extracted during the previous frame denoising, to refine the final noise levels that would be involved in the Wiener filtering. An ExpMA iterative procedure has been selected for the noise estimation

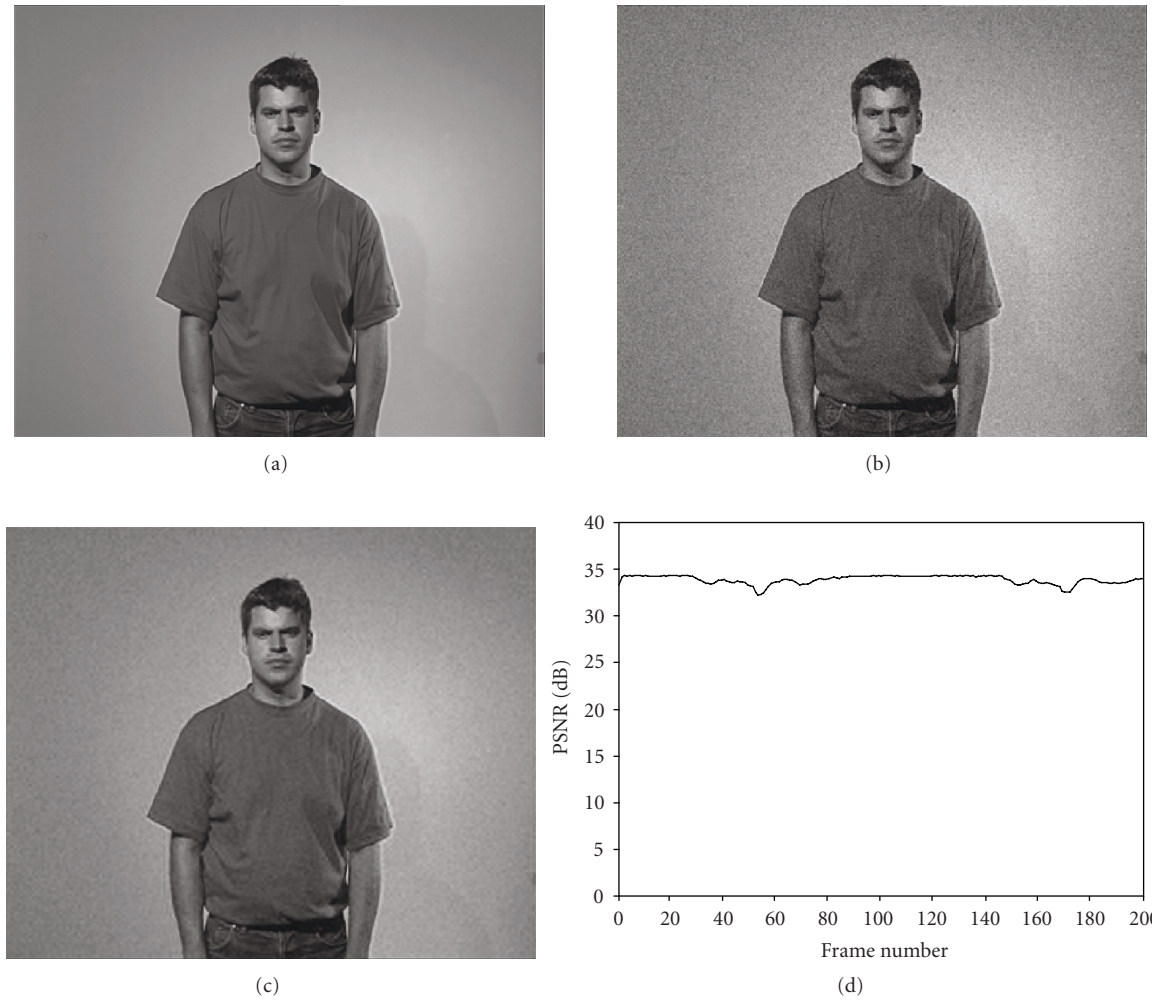


Figure 6: Quantitative analysis of denoising results: (a) original (noise-free) video frame, (b) noise-contaminated image, (c) JWVD-MAD denoised frame, (d) PSNR curves.

process, since it proved very efficient in 1D processing [30], as well as because the whole motion detection process utilizes ExpMA structures:

$$|J_{N\sim 2}(w_i, w_j, n) = a_N \cdot J_{N\sim 1}(w_i, w_j, n) + (1 - a_N) \cdot J_{N\sim 4}(w_i, w_j, n-1), \tag{9}$$

where a_N is the corresponding ExpMA constant ($0 < a_N < 1$), also called memory term [30], $J_{N\sim 4}(w_i, w_j, n-1)$ is the previous-frame noise estimation (extracted after the $(n-1)$-frame denoising has been completed) and $J_{N\sim 1}(w_i, w_j, n)$ is the noise extracted during the first-level denoising of the empirical Wiener filter. The factor k_m might be different at various scales, so we use the generic expression k_m for all $(l; \mathrm{AD})|_{\mathrm{DWT}}$. In fact, we selected to use a unique multiplicative factor for all the detail coefficients k_m for all $(l; \mathrm{AD})|_{\mathrm{DWT}} \neq (Lw; \mathrm{LL})$, except from the $k_m^{(Lw;\mathrm{LL})}$ factor that was adopted for the approximation subimage:

$$\begin{aligned} J_{N\sim 1}(w_i, w_j, n) &= J_X(w_i, w_j, n) - J_{S\sim 1}(w_i, w_j, n) \\ J_{S\sim 1}(w_i, w_j, n) &= \frac{J_X(w_i, w_j, n)}{|J_X(w_i, w_j, n)|} \cdot \max\{[\,|J_X(w_i, w_j, n)| - \mathrm{Th}_N], 0\} \\ \mathrm{Th}_N &= k_m \cdot \sigma_N, \\ \sigma_N &= \frac{\mathrm{Median}(J_X^{(1;\mathrm{HH})}(w_{1i}, w_{1j}, n))}{0.6745}, \quad \forall (l; \mathrm{AD})|_{\mathrm{DWT}} \end{aligned} \tag{10}$$

The refined noise estimation $J_{N\sim 2}(w_i, w_j, n)$ is then introduced to the parametric wavelet Wiener filter (3) and the *WD-EWF* is completed providing the new estimations for

signal and noise wavelet coefficients:

$$J_{S\sim 2}(w_i,w_j,n)=\begin{cases}\left[1-c_{\mathrm{WF}}\cdot\dfrac{P_{N\sim 2}(w_i,w_j,n)}{P_X(w_i,w_j,n)}\right]\cdot J_X(w_i,w_j,n),\\ \qquad \text{if } c_{\mathrm{WF}}\cdot\dfrac{P_{N\sim 2}(w_i,w_j,n)}{P_X(w_i,w_j,n)}\le 1,\\ 0,\quad \text{otherwise}\end{cases}$$

$$J_{N\sim 3}(w_i,w_j,n)=J_X(w_i,w_j,n)-J_{S\sim 2}(w_i,w_j,n),\quad \forall(l;\mathrm{AD})\,|_{\mathrm{DWT}}. \tag{11}$$

The motion detection procedure is then applied using the noise-free coefficients $J_{S\sim 2}(w_i,w_j,n)$ and the $(n-1)$-frame coefficients $J_{S\sim 3}(w_i,w_j,n-1)$, extracted from the complete spatiotemporal filtering in the exact previous step (the refined motion-detection equations are analyzed in the next paragraph). A final task is the implementation of temporal filtering to take advantage of the image similarities between successive frames (especially at motionless locations). Thus, iterative temporal smoothing is employed via a "weighted" ExpMA procedure. Subband moving point matrices $M_{\mathrm{WP}}(w_i,w_j,n)$, provided by motion detection analysis as follows in (14) are utilized to avoid blurring at motion edges:

$$J_{S\sim 3}(w_i,w_j,n)=\begin{cases}a_{\mathrm{TF}}\cdot J_{S\sim 2}(w_i,w_j,n)+(1-a_{\mathrm{TF}})\cdot J_{S\sim 3}(w_i,w_j,n),\\ \qquad \text{if } M_{\mathrm{WP}}^{(l;\mathrm{AD})}(w_i,w_j,n)=0,\\ \qquad \forall(l;\mathrm{AD})|_{\mathrm{DWT}}\\ J_{S\sim 2}(w_i,w_j,n),\quad \text{otherwise,}\end{cases} \tag{12}$$

where a_{TF} is the "temporal filtering" constant of the corresponding ExpMA procedure. The above settlement is quite common to many temporal-filtering-based video denoising algorithms [17, 37], with various modifications encountered ccording to the involved motion detection/estimation parameters. The noise estimations are also refined following the outcome of (12) and the $J_{N\sim 4}(w_i,w_j,n)$ components are extracted similarly to the $\mathbf{J}_{N\sim 1}$ and $\mathbf{J}_{N\sim 3}$ matrices (10), (11). Both $J_{S\sim 3}(w_i,w_j,n)$ and $J_{N\sim 4}(w_i,w_j,n)$ signals would be further utilized at the next iteration (processing at $(n+1)$ frame).

4.2. Dynamic background-foreground segmentation for video motion activity analysis

Having estimated the noise-free signal components $\mathbf{J}_{S\sim 2}(n)$ and $\mathbf{J}_{S\sim 3}(n-1)$, the motion-activity-detection task is performed using the wavelet-adapted ExpMA procedures, suggested by Töreyin et al. [26]:

$$D(w_i,w_j,n+1)=\begin{cases}a_m\cdot D(w_i,w_j,n)+(1-a_m)\cdot J_{S\sim 2}(w_i,w_j,n),\\ \qquad \text{if } (w_i,w_j,n) \text{ is moving}\\ D(w_i,w_j,n),\quad \text{otherwise}\end{cases}\quad \forall(l;\mathrm{AD})\,|_{\mathrm{DWT}}, \tag{13}$$

$$M_{\mathrm{WP}}(w_i,w_j,n)=|J_{S\sim 2}(w_i,w_j,n)-J_{S\sim 3}(w_i,w_j,n-1)|>T_W(w_i,w_j,n),\quad \forall(l;\mathrm{AD})\,|_{\mathrm{DWT}}, \tag{14}$$

$$T_W(w_i,w_j,n+1)=\begin{cases}a_m\cdot T_W(w_i,w_j,n)\\ +(1-a_m)\cdot c_m\cdot|J_{S\sim 2}(w_i,w_j,n)-D(w_i,w_j,n)|,\\ \qquad \text{if } (w_i,w_j,n) \text{ is moving}\\ T_W(w_i,w_j,n),\quad \text{otherwise}\end{cases}\quad \forall(l;\mathrm{AD})\,|_{\mathrm{DWT}} \tag{15}$$

$$M_{\mathrm{WB}}(w_i,w_j,n)=|J_{S\sim 2}(w_i,w_j,n)-D(w_i,w_j,n)|>T_W(w_i,w_j,n),\quad \forall(l;\mathrm{AD})\,|_{\mathrm{DWT}}. \tag{16}$$

The "wavelet motion subimages" $M_{\mathrm{WB}}(w_i,w_j,n)$ are computed according to the original methodology (7), by comparing intensity coefficients with estimated backgrounds (16). However, there are two basic novelties that are introduced in the proposed algorithm, in order to face the noise-caused problems, as well as to satisfy the dynamic BRFR demands, previously mentioned. As already stated, the presence of noise, leads to the erroneous detection of many "isolated moving pixels". Besides denoising, we decided to incorporate "structural decision rules" similar to those proposed for video denoising [15, 18]. Specifically, a moving point (w_{li},w_{lj},n) is considered as "valid movement", only if it belongs to a broader moving region (structure/object); if not, it must be indicated as "false movement" caused by the noise originated differences. In other words, there have to be an adequate number of neighboring active (moving) points, referred as supporting points. This rule was primarily proposed for the validation of the moving pixels $M_{\mathrm{WP}}(w_i,w_j,n)$, calculated via (13), and it is applied to all the involved wavelet subimages. Additionally, it was proved to be helpful for the refinement of the motion subimages $M_{\mathrm{WB}}(w_i,w_j,n)$, estimated as the difference between the background and the frame-intensity (16). The "supporting moving point" threshold was configured based on empirical observation and was adjusted to $T_{\mathrm{SMP}}=3$. Once the subimages $M_{\mathrm{WP}}(w_i,w_j,n)$ and $M_{\mathrm{WB}}(w_i,w_j,n)$ are refined, an upscaling is necessary to construct the original motion images $M_P(i,j,n)$ and $M_B(i,j,n)$. We followed the upscale by 2 rules proposed in [26], where each moving point at level l is transformed to $2^l\times 2^l$ area in the original image dimensions.

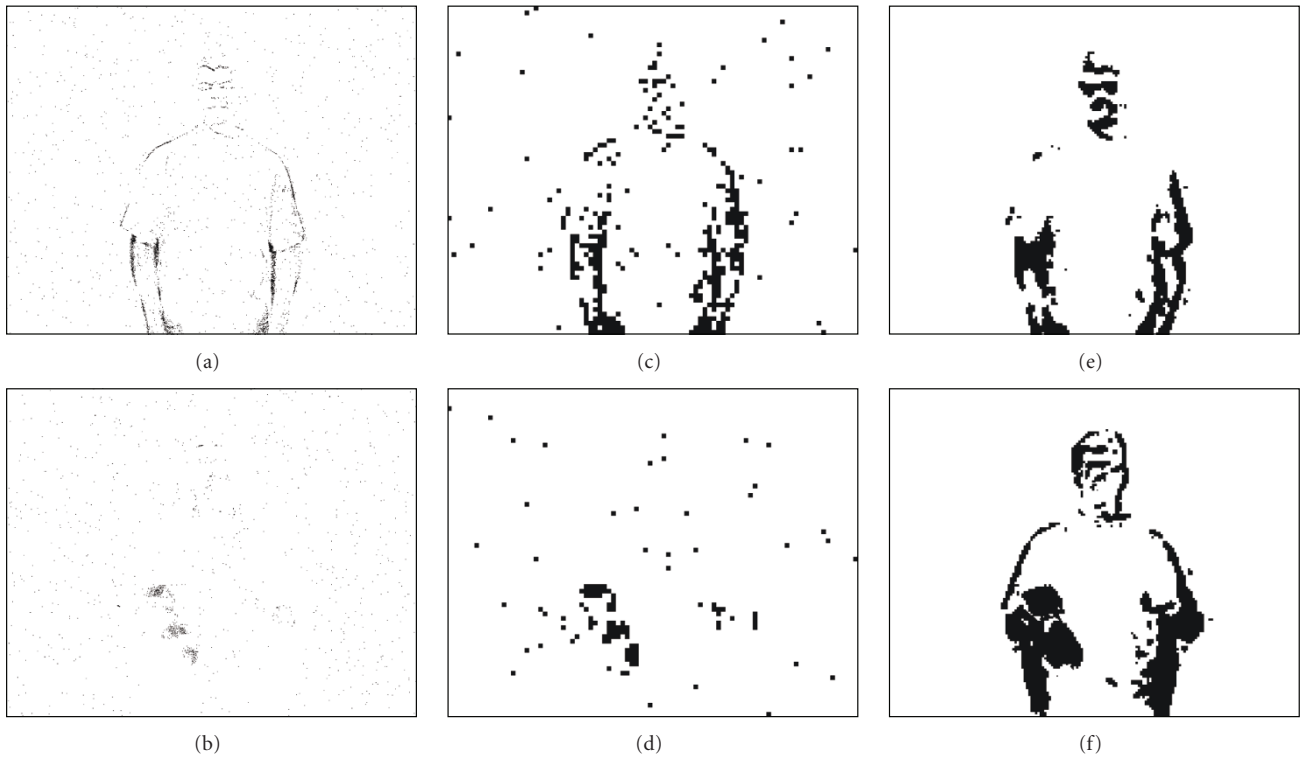


Figure 7: Quantitative analysis of motion detection results: (a)-(b) motion images extracted with the TD-BRFR method, (c)-(d) motion images extracted with the WD-BRFR method, (e)-(f) motion images extracted with the JWVD-MAD algorithm.

This rule can be easily applied for the case of Haar wavelets [26], or for any other mother wavelet, if periodic extension is employed. Alternatively, it is feasible to form all the equivalent motion images and to restrict their dimension to the one of the original image. An additional difference from the WD-BRFR method [26] is that all the involved DWT image coefficients are used (all the detail coefficients plus the approximation coefficients at the lowest level $l = L_W$, in contrast to [26] where only the lowest decomposition level coefficients are used).

The second modification deals with the fact that dynamic BRFR segmentation is necessary. Human activity monitoring has specific particularities when compared to classical video surveillance cases, such as traffic monitoring or security systems. Thus, only a portion of the original background is actually revealed, while parts of the human subjects belong to stationary background for specific periods of time. If a movement occurs, this dynamic background may change, so that it is necessary to reestimate a more appropriate background image. Considering that neither background images nor thresholds are updated when pixels are moving, the simplest solution to the adaptive BRFR task is to reinitiate the WD-BRFR procedure, once a significant movement has been completed. In this way, background is estimated from scratch using the intensities of nonmoving frames. The only unsettled issue is the implementation of a decision system to indicate the restarting operation.

A simple metric to quantify the motion detection is to sum-up all the binary values $M_P(i, j, n)$ or $M_{\mathrm{WP}}(w_i, w_j, n)$, in order to calculate the motion intensity $m_{\mathrm{int}}(n)$, by means of total number of moving points per frame [1]:

$$\begin{aligned} m_{\mathrm{int};P}(n) &= \sum_{i=0}^{N_H-1} \sum_{j=0}^{N_V-1} M_P(i, j, n) \\ &\approx \sum_{(l;\mathrm{AD})} \sum_{w_i} \sum_{w_j} M_{\mathrm{WP}}(w_i, w_j, n), \end{aligned} \tag{17}$$

where the P subscript is used to index that the specific operant applies to the moving pixels array $M_P(i, j, n)$. The B subscript is alternatively used for the motion images $M_B(i, j, n)$. "1D motion signals" can be effectively deployed to facilitate motion-based video summarization and abstraction. It is important to mention that the motion intensity parameter described in (8) is completely different from the "MPEG-7 motion intensity parameter," which has been established via experimental procedures considering perceptual aspects of the human vision [19–21]. To avoid confusion, we will use the "motion equivalent surface" (m_{SE}) index instead, which is equal to the square root of m_{int}. The m_{SE} has the advantage that features smoother changes, and it also has a physical interpretation that is easier to follow showing the "equivalent moving area."

The $m_{\mathrm{SE};P}$ parameter was employed for process reinitiation according to the following basic steps.

(a) Significant event motion is indicated as soon as the $m_{\mathrm{SE};P}(n)$ value exceeds an empirical defined threshold T_{event} (values of T_{event} between 15–50 worked

efficiently in the 720 × 576 images of our application). An additional constrain is that the previous $m_{SE;P}(n-1)$ value should be lower than the present.

(b) A Boolean flag FL is activated once a significant motion event is detected.

(c) When the $m_{SE;P}(n)$ falls below the threshold (and it is in decreasing order: $m_{SE;P}(n-1) > m_{SE;P}(n)$), the motion event completes and the *WD-D-BRFR* algorithm reinitiates. The flag FL is also deactivated for future events detection.

(d) Finally, time constrains are introduced to automatically reinitiate the *WD-D-BRFR* process if the FL parameter remains idle for a long period of time (i.e., >200 frames).

Thus, the detection of a new video event (v_E) at frame n and the reinitiation decisions are updated in combination with the FL sequence according to the following Boolean formulas:

$$\begin{aligned}
\mathrm{FL}_{\mathrm{ON}}(n) &= \{\overline{\mathrm{FL}(n)}\} \text{ AND } \{m_{SE;P}(n) > T_{\mathrm{event}}\} \\
&\quad \text{AND } \{m_{SE;P}(n) > m_{SE;P}(n-1)\} \\
\mathrm{FL}_{\mathrm{OFF}}(n) &= \{\mathrm{FL}(n)\} \text{ AND } \{m_{SE;P}(n) < T_{\mathrm{event}}\} \\
&\quad \text{AND } \{m_{SE;P}(n) < m_{SE;P}(n-1)\} \\
\mathrm{FL}(n+1) &= \langle \{\overline{\mathrm{FL}(n)}\} \text{ AND } \{\mathrm{FL}_{\mathrm{ON}}(n)\} \rangle \\
&\quad \text{OR } \langle \{\mathrm{FL}(n)\} \text{ AND } \{\mathrm{FL}_{\mathrm{OFF}}(n)\} \rangle,
\end{aligned} \tag{18}$$

where the $\mathrm{FL}_{\mathrm{ON}}/\mathrm{FL}_{\mathrm{OFF}}$ parametes indicates the detection/completion of a new video event, respectively, while the $\mathrm{FL}_{\mathrm{OFF}}$ condition also triggers process reinitiation. However, the estimation of the exact start-stop timing information needs further refinement (the corresponding analysis is presented in the next paragraph). Considering that background/threshold updates are suspended when moving pixels are detected, it is easy to understand that the WT-D-BRFR reinitiation does not cause instability or similar other problems to the BRFR segmentation procedure.

4.3. Multimodal event detection, segmentation, and summarization (MEDSS)

The outcomes of the JWVD-MAD algorithm are further utilized as inputs to a "multimodal event detection segmentation and summarization" (MEDSS) methodology, to facilitate content indexing and abstraction. Specifically, the extracted motion parameters $m_{SE;P}(n)/m_{SE;B}(n)$ and the flag sequences FL(n) are fed to a video event detection, segmentation, and summarization (VDSS) system. VDSS determines the total number (NV_E) of the detected video events v_E and their exact starting ($v_{E;\mathrm{IN}}$)/ending ($v_{E;\mathrm{OUT}}$) locations. In addition, sound processing is performed to all the audio-surveillance and bioacoustic recordings available. In this way, "automated audio-detection segmentation and indexing" (AADSI) is conducted in order to estimate the corresponding sound and bioacoustic events (s_E and b_E, resp.). A counterpart AADSI methodology has been developed, taking advantage of the multiresolution scanning approach of the long-term wavelet-based detection, segmentation, and summarization (LT-WDSS) algorithm [1, 12, 38].

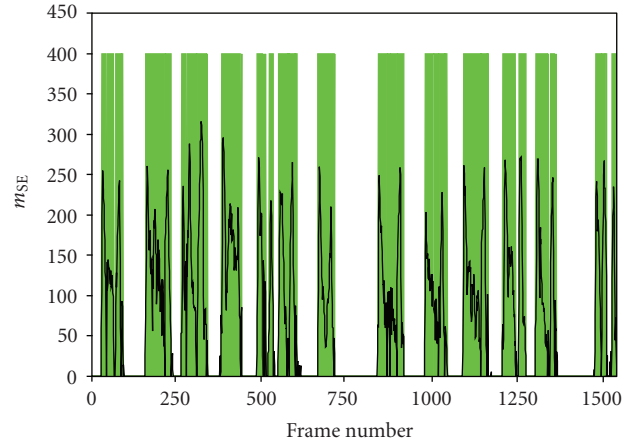


Figure 8: Motion activity curve (black curve) and video motion detection results (green color) automatically extracted via the VDSS method ($T_{\mathrm{event}} = 40$): experimental procedure with artificial noise-contamination ($\sigma_N^2 = 100$) using sign-language videos.

Besides the determination of the sound/bioacoustics events, energy-comparisons between the tracks of the multichannel recordings are performed for topographic analysis purposes, while spectrographic colormaps and power envelope curves are employed for summarization purposes [1, 12, 38]. Since the AADSI methodology is well presented in the related [1, 12, 38], we will focus our attention to the VDSS method, as well as the interaction between the three content types (video, sound, and bioacoustic events).

It is clear that the motion intensity sequence $m_{SE;B}(n)$ provides an overview of the video motion changes via 1D plots. In addition, the flag on/off timing estimated during the WD-D-BRFR process is useful in detecting video events. Specifically, the flag-on/flag-off points are extended until the $m_{SE;B}(n)$ curves meet a local minimum, so that a video event is localized (still frame overheads might be also included),

$$\begin{aligned}
v_{E;\mathrm{IN}} &= n : \min\{m_{SE;B}(n) : m_{SE;B}(n-1) \geq m_{SE;B}(n) \\
&\qquad < m_{SE;B}(n+1)\}, \quad n \leq \arg_n(\mathrm{FL}_{\mathrm{ON}}) \\
v_{E;\mathrm{OUT}} &= n : \min\{m_{SE;B}(n) : m_{SE;B}(n-1) > m_{SE;B}(n) \\
&\qquad \leq m_{SE;B}(n+1)\}, \quad n \geq \arg_n(\mathrm{FL}_{\mathrm{OUT}}).
\end{aligned} \tag{19}$$

Optionally, the energy of the "inside flags" $m_{SE;B}(n)$ may also be compared with predefined thresholds, in order to avoid registering many small and random movements as significant events. Similarly, two or more detected events in row may be concatenated (based on their temporal distance and the demands of each application), avoiding unnecessary splitting of self-contained video episodes.

After video event detection has been completed, highlighting images (HLI) are also extracted for video summarization purposes. We have decided to extract 3 highlighting frames for every video episode. The first $\mathrm{HLI}_{\mathrm{IN}}(v_E)$ and last frames $\mathrm{HLI}_{\mathrm{OUT}}(v_E)$ of each detected event provide image instances just before and after the specific episode. The internal

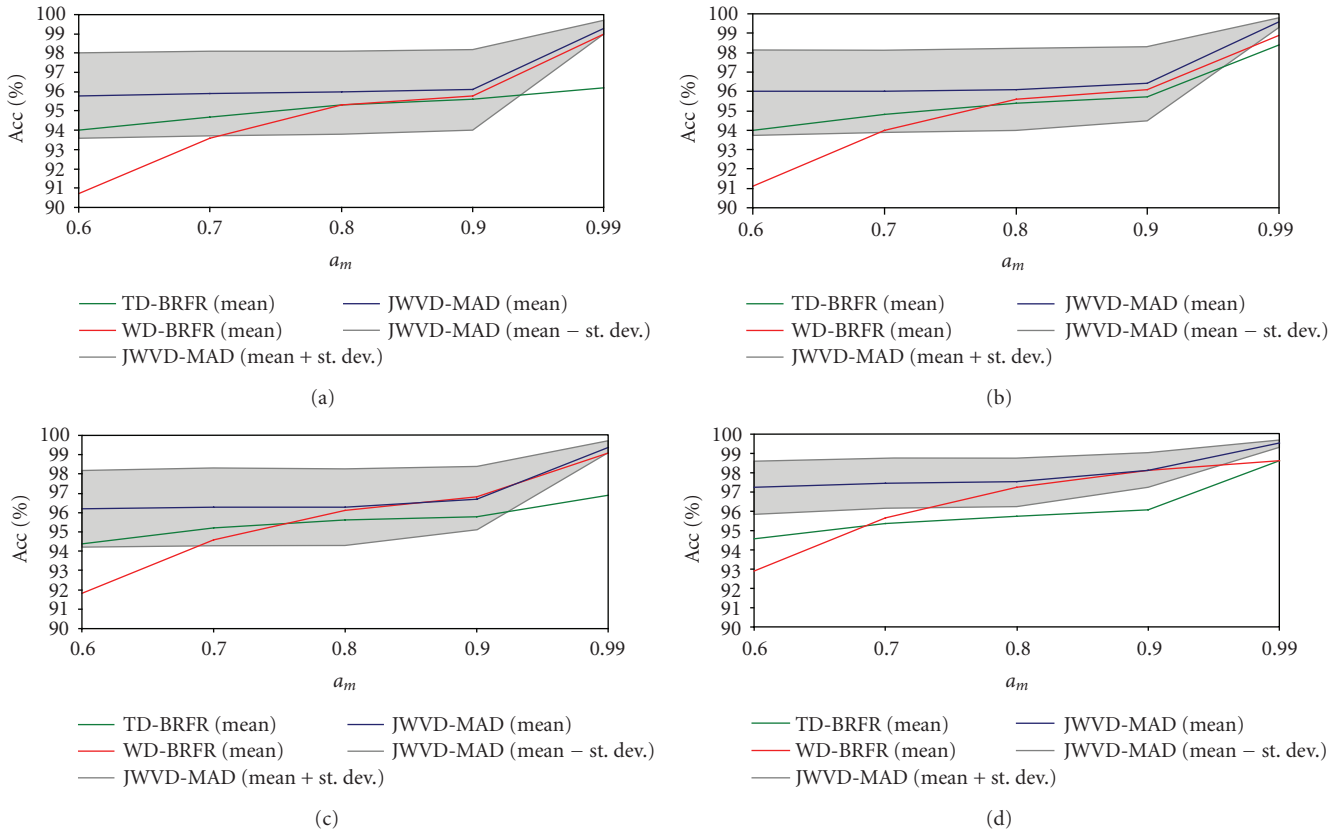


Figure 9: Sensitivity analysis of the three BRFR segmentation methods (TD-BRFR, WD-BRFR, JWVD-MAD), using computer generated image sequences (a box graphic scrolling diagonally across the scene). The mean values (plus/minus the standard deviation for the JWVD-MAD case) of the pixel-based accuracy (Acc) metric versus the a_m parameter are plotted (the remaining BRFR parameters are set to $c_m = 5$; $T_{\text{SMP}} = 1$, for the JWVD-MAD) for four different noise-contaminated test videos: (a) slow passing low contrast (SPLC), (b) slow passing high contrast (SPHC), (c) fast passing low contrast (FPLC), (d) fast passing high contrast (FPHC).

frame $\text{HLI}_{\text{INT}}(v_E)$ that features the highest motion activity is additionally selected, in order to be able to synopsize the "action" of the episode:

$$\begin{aligned}
\text{HLI}_{\text{IN}}(v_E) &= \text{IDWT}_{2D}\{J_{S\sim 3}(w_i, w_j, n_{\text{IN}})\}, \\
n_{\text{IN}} &= \arg_n(v_{E;\text{IN}}) \\
\text{HLI}_{\text{INT}}(v_E) &= \text{IDWT}_{2D}\{J_{S\sim 3}(w_i, w_j, n_{\text{INT}})\}, \\
n_{\text{INT}} &= \arg\max_{n_{\text{IN}} \le n \le n_{\text{OUT}}}(m_{\text{SE};B}) \\
\text{HLI}_{\text{OUT}}(v_E) &= \text{IDWT}_{2D}\{J_{S\sim 3}(w_i, w_j, n_{\text{OUT}})\}, \\
n_{\text{OUT}} &= \arg_n(v_{E;\text{OUT}})
\end{aligned} \tag{20}$$

avoiding to execute inverse 2D-DWT (IDWT_{2D}), except for the highlighting images cases. A full set of binary motion images $\{\text{BMI}\} = \{M_B(i, j, n)\}$ is also extracted for every one of the detected events to synopsize the "human activity," offering the advantage of fast browsing and easy manipulation due to the 1-bit resolution (binary arrays).

Another important issue concerns multimodal interaction between the three content types, namely, video surveillance, sound surveillance, and bioacoustic monitoring. These sound, video, and bioacoustic events might occupy complete different time periods of the experimental/surveillance procedure, but there are also many cases that events are activated simultaneously for more than one content type, so that a multimedia event m_E is formed. The Boolean expression suggesting the registration of a multimedia event m_E is given in the formula below, while refinement of the event's timing is also necessary:

$$\begin{aligned}
m_E &: \exists \mu, \lambda = \{v \longleftrightarrow v_E, s \longleftrightarrow s_E, b \longleftrightarrow b_E,\} \Longleftrightarrow [t_{\mu;\text{IN}}, t_{\mu;\text{OUT}}] \\
&\quad \cap [t_{\lambda;\text{IN}}, t_{\lambda;\text{OUT}}] \neq \emptyset, \\
t_{\text{IN}}(m_E) &= \min\{t_{V;\text{IN}}(v_E), t_{S;\text{IN}}(s_E), t_{B;\text{IN}}(b_E)\} \\
t_{\text{OUT}}(m_E) &= \max\{t_{V;\text{OUT}}(v_E), t_{S;\text{OUT}}(s_E), t_{B;\text{OUT}}(b_E)\},
\end{aligned} \tag{21}$$

where $t_{\text{IN}}/t_{\text{OUT}}$ is the time-equivalent starting/ending location of the multimedia event m_E, $t_{V;\text{IN}}/t_{V;\text{OUT}}$, $t_{S;\text{IN}}/t_{S;\text{OUT}}$ and $t_{B;\text{IN}}/t_{B;\text{OUT}}$ are the corresponding timing (start/end locations) of the coincident video v_E, sound s_E, and bioacoustic b_E events.

In the case of multimedia events, further multimodal analysis is enabled to facilitate the long-term inspection process. As stated, bioacoustic events provide information about the human behavior (gastrointestinal motility in our case [12]), such as activity presence or absence,

energy/frequency/duration characteristics, and so forth, that could be further utilized for diagnostic purposes [1, 12–14]. However, misclassified bioacoustic b_E events might be registered due to the presence of human body movements and sliding noises, as well as due to intense dialogues between the subjects and the nursing/medical staff. It is clear that movement artefacts can be more easily recognized by combining audio and video detection results [1, 12, 13]. Similarly, energy-based comparisons and cross-correlation metrics between s_E and b_E would allow to detect the presence of ambient noise or any other sound sources that could affect the integrity of the bioacoustic recordings or even the human psychophysiological response. For instance, this modality would help to decide whether a strong bioacoustic signal has been recorded from the surveillance mics, or interference to the bioacoustic acquisition system has been occurred due to intense ambient noise [1, 12, 13]. Additionally, the usefulness of the video surveillance information, such as human body position, degree of anxiety, cough, apnoea, and other visual indicated signs, is also related to the evaluation or even the assisted diagnosis of various pathophysiological factors connected with abdominal, cardiac, and lung sounds (related examples and references have been provided in Section 1). A characteristic example, where the proposed methodology is currently used, is the evolution of the "human response to noise" study [39], where (a) audio monitoring provides useful information about the experimental conditions, (b) bioacoustic recordings (such as heart, respiratory, and abdominal sounds) are used as measures to evaluate human psychophysiological response, while (c) video surveillance permit continuous monitoring of the experimental conditions, as well as the human reaction be means of body movements and facial expressions.

The multimodal analysis results are further utilized to extract textual comments and structural annotation (e.g., validation of audio pattern classification results [1, 13], alarm indicators related with the integrity of the acquired data, motion activity rates characterizing human behavior, interpretation of human anxiety, etc.). For example, if coincidence of all audio, video, and bioacoustic events is observed at a specific time-instance, this is likely to be connected with intensive movements and sliding noises. Similarly, if intensive ambient noise is present, events will be detected for the audio monitoring signals, while the initiation of uncorrelated bioacoustic events and the detection of surveillance video events, would probably indicate human controlled reaction or sympathetic arousal. In the case that only "small" video events are detected, a sensible interpretation would be that small human body motion are observed without generating sliding noise (e.g., head/face movements), while the presence of bioacoustic-only events ensures the validity of the acquired biomedical data. Besides the above marginal conditions, intermediate states are more often observed, where various combinations of the three signal entities are encountered in different intensities and duration/repetition cycles. The incorporation of expert systems [1, 13] as well as various other tools for content characterization, semantic annotation and structural classification, and their integration to all the three sources of audio-visual information can be very helpful towards efficient content description and management [13]. In fact, the data structures with the semiautomated extracted information of the MEDSS approach are currently employed to train more sophisticated pattern recognition systems for content classification and characterization. In any case, we have decided to use different data structures and files to store the content description information, than to incorporate them to the original recordings, following the "bits about the bits" philosophy of the MPEG-7 protocol [1, 19–21]. A related work, where the multimodal content interaction and the MPEG-7 schemas, that are employed to hold content descriptions, are currently under preparation.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed methodology was tested on video-assisted bioacoustic monitoring applications, aiming to provide new potentials in noninvasive diagnosis of gastrointestinal motility dysfunctions [1, 12–14]. The recordings took place on the premises of the Papageorgiou General District Hospital in Thessaloniki. Semiprofessional digital-8 camcorders where used, allowing video data transfer directly in digital format to a PC via DV protocol (IEEE-1394). Thus, video sequences were coded as DVPAL files with resolution 720 × 576 ($N_H = 576$, $N_V = 720$). A dual camera system was used, providing wide and zoom view of the subjects under bioacoustic monitoring, while night vision was engaged during overnight recordings, which was selected in the majority of the experiments [1]. As already stated, color discarding was decided during preprocessing for homogeneity purposes (between diurnal and nocturnal recordings), as well as color information is not necessary for both the automated and manual inspection processes. A dual-microphone sound surveillance system was employed, with use of the cameras' mics [1]. A seven-channel human bioacoustic monitoring system was also engaged [1, 12]. All the implementations were developed in the LabVIEW 7.1^{TM} software environment, using the add-on signal processing toolset in combination with "avi"/IMAQ-vision libraries.

5.1. Qualitative analysis

Original recordings with durations ranging between one and six hours were employed throughout the setup and calibration process of the developed JWVD-MAD method and are also used for the qualitative analysis that follows. Based on empirical observation, as well as on quantitative validation procedures described, the JWVD-MAD method was implemented selecting 2-level ($L_W = 2$) DWT and it was adjusted using the parameters $c_{\mathrm{WF}} = 2.5$, $a_m = 0.99$, $c_m = 6$, $T_0 = 50$, $k_m^{(l;\mathrm{AD})} = 1.5$ (except from $k_m^{(lw;\mathrm{LL})} = 0.15$), $a_N = 0.95$, and $a_{\mathrm{TF}} = 0.8$. Besides empirical observations on natural biomedical video recordings, various validation procedures were implemented for the final adjustment of the above parameters. Thus, good quality video sequences, featuring similar technical characteristics with our content (720 × 576, DVPAL) was selected and artificially contaminated with noise, in order to be used for method evaluation. Specifically, sign-language videos were selected, since they are also

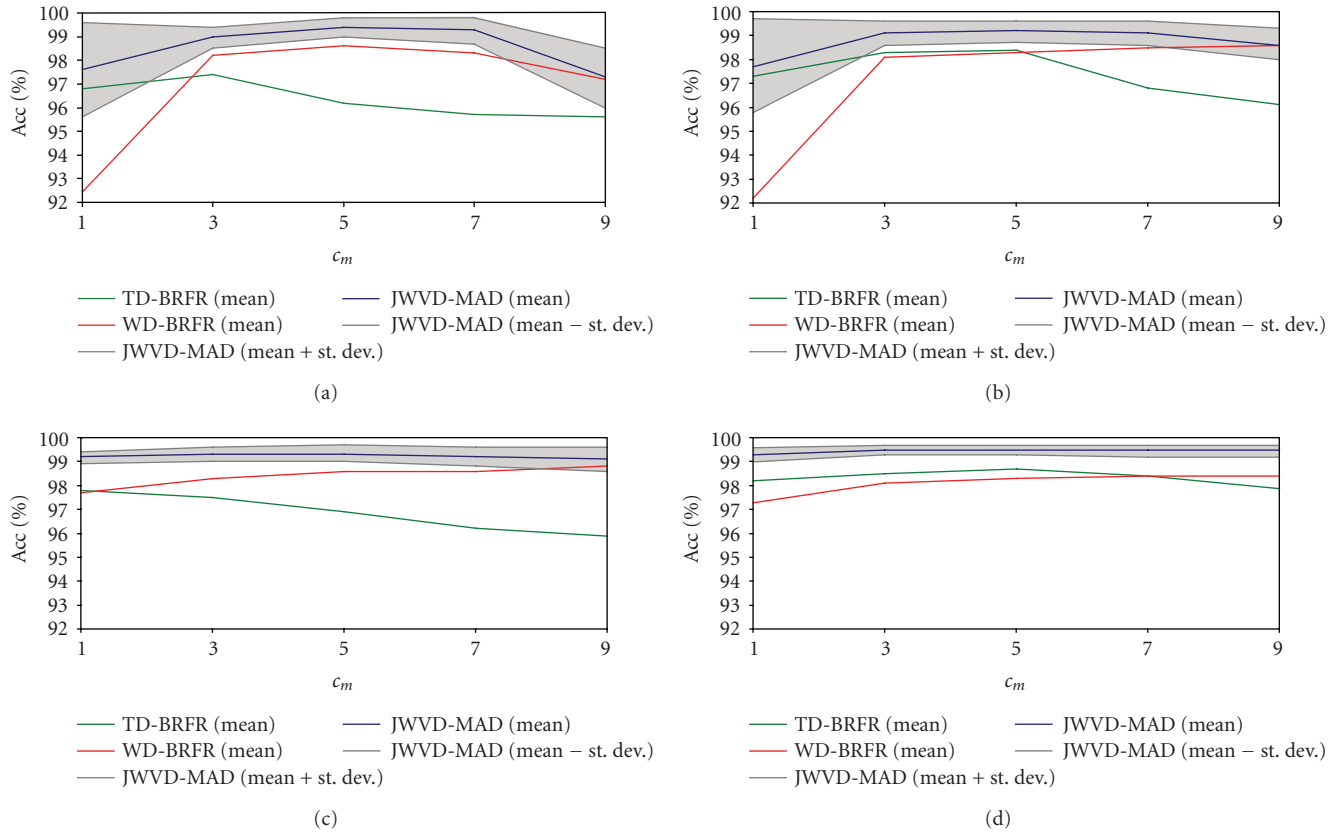


FIGURE 10: Sensitivity analysis of the three BRFR segmentation methods (TD-BRFR, WD-BRFR, JWVD-MAD), using computer generated image sequences (a box graphic scrolling diagonally across the scene). The mean values (plus/minus the standard deviation for the JWVD-MAD case) of the pixel-based accuracy (Acc) metric versus the c_m parameter are plotted (the remaining BRFR parameters are set to $a_m = 0.95$; $T_{\text{SMP}} = 1$, for the JWVD-MAD) for four different noise-contaminated test videos: (a) slow passing low contrast (SPLC), (b) slow passing high contrast (SPHC), (c) fast passing low contrast (FPLC), (d) fast passing high contrast (FPHC).

describing human activity where the proposed methodology could be applied to facilitate motion detection, segmentation, and content indexing. Most of the parameters related to video denoising where coordinated based on the peak signal-to-noise ratio (PSNR) [14, 15, 31], under the presence of Gaussian additive noise with known characteristics (noise variances σ_N^2 between 100–200 were tested).

As already stated, qualitative analysis was based on the available human surveillance recordings, and it was very helpful during the entire setup of the method. In general, we had to evaluate three different aspects: video denoising performance, motion detection efficiency, and event-detection accuracy. Figure 2 presents two denoising examples with severe noise contamination problem. Based on these results we may claim that video denoising is quite satisfactory under these extreme conditions (the noise variance was estimated quite above 100). Even more important is the fact that motion detection results were quite satisfactory under these circumstances. Figure 3 provides comparisons between the motion images extracted with the proposed JWVD-MAD approach and the TD-BRFR, WD-BRFR methods proposed in [25, 26], respectively. These examples, and all the motion detection comparisons that follow, were extracted using the reference methods to small-time periods, without the necessity to reinitiate process, since they do not meet the dynamic BRFR conditions already discussed. Joint evaluation with the baseline algorithms were conducted for two reasons: (a) in order to demonstrate the improvements made with the new methodology, and (b) for performance comparisons, since they all are general algorithms proposed for video motion detection. Returning to the analysis of Figure 3, it is obvious that the erroneous estimated moving pixels (randomly distributed-isolated dots) are by far less in our approach, than in the other two methods. The denoising procedure as well as the structural motion detection aspects and the "supporting neighboring pixel" hypothesis are the basic reason of the observed improvements. These above results are also explained from the motion intensity curves in Figure 4. It is obvious that JWVD-MAD curves are by far less noisy in such conditions (a noise variance around $\sigma_N^2 = 150$ was estimated), where even the smooth, manually tagged, "head-turn" event is detected as significant activity. Although the experiments were conducted to short-term recordings, the motion curves provided by the TD-BRFR and WD-BRFR techniques seem to be more and more random as the frame number increases, issue that validates the dynamic process initiation procedure that was followed in our case. Another motion-activity curve is presented in Figure 5, along with the

video detection/segmentation "flagging." It is obvious that the proposed VDSS analysis is very helpful in detecting human activity and movement artifacts in long-term monitoring periods.

5.2. Quantitative analysis

Quantitative analysis was performed on the basis of noiseless video recording that was artificially noise-contaminated for comparison purposes. Greek sign-language videos acquired at ideal TV studio conditions were used for this purpose. Figure 6 provides comparison between initial, noised, and JWVD-MAD reconstructed video frames that were contaminated with additive Gaussian noise ($\sigma_N = 15$). The PSNR parameter was estimated near to 35 dB, which is a quite good result for the current noising conditions. Figure 7 presents the estimated motion images for the given noise conditions. It is obvious that the method manages to effectively detect motion regions, successfully suppressing the noise effects. Figure 8 provides video detection/segmentation results based on the related VDSS methodology. Closely spaced events can be successfully isolated even with the presence of significant noise, as long as the appropriate timing parameters are selected to determine the desired resolution. Besides these examples, various subjective tests for the evaluation of the detection/segmentation efficiency were performed, resulting in an efficiency of above 90% of both the number and the location of correctly located events (considering that a distance less than two seconds between the manual and automated starting/ending points is classified as a true positive detection). Specifically, various sign-language words were selected and randomly distributed to different time locations, in video recordings that were contaminated with noise. Students of the Laboratory of Electroacoustics and TV Systems were engaged to manually locate the video episodes inside limited duration recordings (i.e., 5 minutes). The location results and the related number of detected events were compared to the automated results of the JWVD-MAD algorithm and the VDSS method. Finally, denoising results and comparisons with standard video sequences and classical denoising approaches were also tried.

5.3. Influence of the JWVD-MAD parameters and sensitivity analysis

The difficulty that someone might face when using the JWVD-MAD algorithm and the subsequent VDSS technique is connected to the fact that many parameters have to be configured manually. However, a more careful examination would reveal that the parametric nature of both the previously stated methods tend to offer more advantages rather than disadvantages. For instance, the parameters can be configured for optimal performance according to the demands of a certain application, offering the ability of utilization to many surveillance-related fields, including various human activity analysis approaches. Thus, the only issue that remains unsettled is connected with the procedures that should be followed in order to achieve optimal configuration. As already stated, empirical observation on natural video-surveillance recordings were employed in combination with various metric in order to configure the filter parameters. Additionally, sensitivity analysis was necessary in order to demonstrate the influence of each parameter to the response of the JWVD-MAD algorithm. In general, we may distinguish two types of parameters: (a) the first category includes a_N, $k_m^{(l;\mathrm{AD})}$, c_{WF}, and a_{TF} that control the video denoising process, while (b) the a_m, c_m, T_0, T_{SMP}, and T_{event} parameters are related to the BRFR segmentation and the subsequent motion and video event detection processes. We will focus our attention to the second category, the motion detection parameters, since they play a more significant role to the human activity analysis procedure. As discussed earlier, the denoising parameters were configured with the use of artificially noise-contaminated videos and metrics like the peak signal-to-noise ratio (PSNR) and the mean-square error (MSE) of the restoration process, procedure that is quite common in most signal denoising evaluation approaches [1, 15–18, 30]. A corresponding paper that is particularly focused on the performance and evaluation of the video denoising method is currently under preparation. Thus, the selected values of $c_{\mathrm{WF}} = 2.5$, $k_m^{(l;\mathrm{AD})} = 1.5$ (except from $k_m^{(lw;\mathrm{LL})} = 0.15$), $a_N = 0.95$, and $a_{\mathrm{TF}} = 0.8$ will be considered for the analysis that follows.

Performance evaluation of tracking and surveillance results is very important in most video monitoring/surveillance applications [40]. In those cases, the "ground truth" of the BRFR segmentation is necessary, while various approaches are followed to perform this task. For instance, video data-bases where manual object detection tagging has applied may be used, while comparison with the results of well-accepted video tracking methods is also very common [40]. Another option is to generate synthetic image sequences (computer graphics) where the ground truth of the BRFR segmentation is easily obtained. Although the evaluation process might be biased due to the fact that the BRFR algorithms are tuned to the specific, unrealistic, surveillance content [40], sensitivity analysis is still very useful since it shows how the parameters influence the motion-detection accuracy. Another unsettled issue is related to the choice of the appropriate metric to demonstrate the detection performance. In general we may distinguish the pixel-based and the object-based metrics, where various perceptual tasks are usually involved [40]. Considering the first case, the simplest pixel-based metric is the accuracy (Acc) that expresses the percentage of the correctly classified pixels as moving and nonmoving [40]:

$$\mathrm{Acc} = \frac{N_{\mathrm{tp}} + N_{\mathrm{tn}}}{N_{\mathrm{pixels}}} \cdot 100, \tag{22}$$

where $N_{\mathrm{tp}}/N_{\mathrm{tn}}$ is the number of the correctly classified moving/nonmoving pixels and N_{pixels} the total number of pixels (equals the image resolution product).

Based on the above remarks, we decided to generate a grayscale-gradient rectangular box (188 by 140 pixels) as an object that enters and leaves an empty (black) background scene, using two different speeds (fast and slow passing, FP/SP; 1.5 and 3 seconds duration, resp.). In addition, we

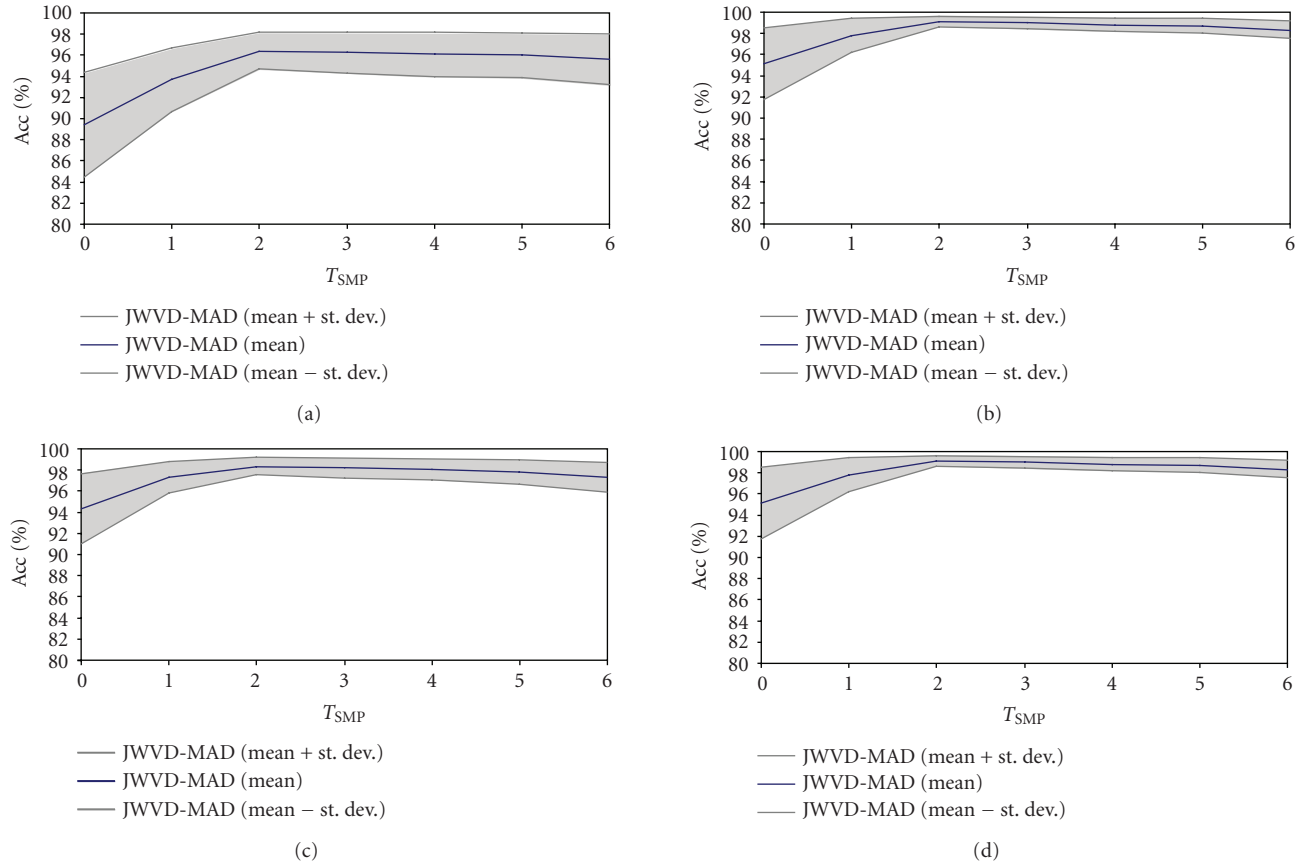


Figure 11: Sensitivity analysis of the JWVD-MAD algorithm, using computer generated image sequences (a box graphic scrolling diagonally across the scene). The mean values (plus/minus the standard deviation) of the pixel-based accuracy (Acc) metric versus the T_{SMP} parameter are plotted (the remaining BRFR parameters are set to $a_m = 0.95$; $c_m = 5$) for four different noise-contaminated test videos: (a) slow passing low contrast (SPLC), (b) slow passing high contrast (SPHC), (c) fast passing low contrast (FPLC), (d) fast passing high contrast (FPHC).

decided to test two different contrast levels for the rectangular object (high and low contrast, HC/LC; their dynamic range ratio equals to 2 : 1), suggesting two different dynamic ranges for the corresponding image sequences. The ground-truth motion images were easily extracted for the combination of the above states, so that four video sequences (SPLC, SPHC, FPLC, FPHC) where used as a basis for the comparisons.

Although sensitivity analysis was not performed in the original works of Collins et al. [25] and Töreyin et al. [26], the role of parameters a_m and c_m is clear: the first controls the pace of the adaptation speed by means of the averaging length (background image refinement, thresholds update, etc.), while the second controls the detection sensitivity. Nevertheless, we decided to involve these two parameters to our sensitivity analysis for the sake of completeness. The behavior of these two parameters was one of the reasons that we decided to construct the four test videos (SPLC, SPHC, FPLC, FPHC), with the previously mentioned characteristics. The nature of the synthesized videos has many similarities with the traffic surveillance sequences used in the original works [25, 26], so that there is no need for dynamic BRFR segmentation. Thus, the use of the parameter T_{event} does not apply in the current experiment. In addition the TD-BRFR and WD-BRFR approaches of the original works [25, 26] can be used without the necessity for process reinitiation. Having these remarks in mind, motion detection accuracy was estimated for all the three methods (TD-BRFR, WD-BRFR, and JWVD-MAD), with various values of the parameters a_m, c_m, T_{SMP} employing artificial noise-contamination ($\sigma_N^2 = 100$) to all the four test sequences (SPLC, SPHC, FPLC, FPHC). The sensitivity analysis results are presented in Figures 9 to 11.

We may observe that all the three BRFR methods tend to give better results as the parameter a_m tends to 1 (Figure 9) case that corresponds to longer averaging. Based on Figure 10, the c_m parameter has different effect for each method: small values tend to give better results for the TD-BRFR case, while larger values work better in the WD-BRFR approach. Values between $c_m = 4$ to $c_m = 6$ were proven most suitable for the JWVD-MAD method. In general, accuracy tends to be more stable in a broader range of c_m values for our method compared to the baseline algorithms. Figure 11 proves that the incorporation of the T_{SMP} parameter provides a significant improvement of the detection, with values of T_{SMP} between 2 and 3 giving the best results. It is also obvious that the JWVD-MAD accuracy is enhanced compared to the baseline TD-BRFR and WD-BRFR methods. Although

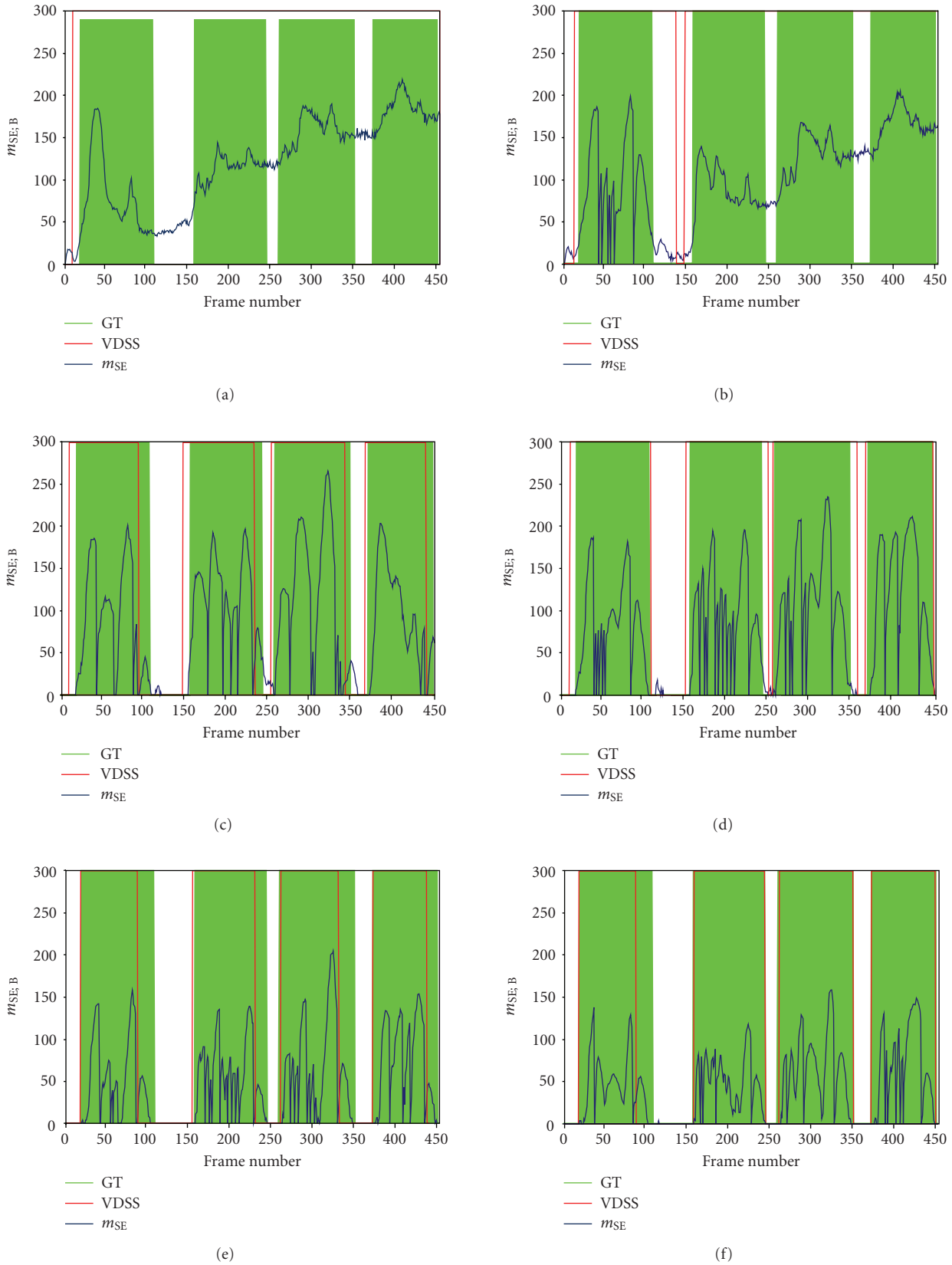


FIGURE 12: Motion-based video detection results using the VDSS method with various parameters of T_{event}, T_{SMP} (the remaining JWVD-MAD parameters were adjusted to $a_m = 0.99$, $c_m = 6$); (the blue curve presents the $m_{SE;B}$ parameter, the red the automated detection/segmentation results and the green indicates the ground truth-GT results): (a) $T_{event} = 20$, $T_{SMP} = 0$, (b) $T_{event} = 50$, $T_{SMP} = 0$, (c) $T_{event} = 20$, $T_{SMP} = 3$, (d) $T_{event} = 50$, $T_{SMP} = 3$, (e) $T_{event} = 20$, $T_{SMP} = 6$, (f) $T_{event} = 50$, $T_{SMP} = 6$.

the improvements in accuracy seems very small (∼1%), it is important to understand that this percentage quantity corresponds to 4147 classified (or misclassified) pixels, which is almost equal to one quarter of the object surface.

In addition to the above results, a second sensitivity analysis procedure was also necessary in order to monitor the influence of the JWVD-MAD parameters to the video detection and segmentation process. Artificially noise-contaminated Greek sign-language videos were again used (because of their closer resemblance to our natural recording compared to the synthesized videos, as well as the ability to fully control noise contamination properties). The motion detection and segmentation ground truth was obtained via manual tagging to the initial, noise-free image sequences. Figure 12 presents the motion detection results for various values of the T_{SMP} and T_{event} parameters together with the manual segmentation of the test image sequences. We may observe that T_{SMP} plays a very significant role in the detection procedure, since erroneous motion estimation may lead to misdetection and wrong segmentation results (Figures 12(a), 12(b)). A possible option to avoid the estimation of exaggerated motion would be to further increase the values of T_{SMP} parameter (Figures 12(e), 12(f); $T_{\text{SMP}} = 6$). However, this leads to the extraction of erroneous binary motion images. Thus, the best solution is to balance the T_{SMP} parameter (Figures 12(c), 12(d); $T_{\text{SMP}} = 3$), combined with a suitable T_{event} selection. Small T_{event} values lead to quite "jerky" behavior of the motion curves (Figures 12(a), 12(c), 12(e); $T_{\text{event}} = 20$), while values around $T_{\text{event}} = 50$ tend to provide more stable results (Figures 12(b), 12(d), 12(f)). Another issue that needs further discussion is that most of the events are closely spaced, fact that is not quite common in natural biomedical monitoring videos. Within this context, the fast-pace sign-language videos were used in the basis of somehow a worst case scenario. In any case, we may observe that very small distances between the extracted time boundaries of the automated event detection and the ground truth results are produced.

5.4. Conclusion and future work

This paper focuses on the implementation of the "joint wavelet video denoising and motion activity detection" methodology, proposed for video enhancement, event detection, and summarization purposes. The purpose of the JWVD-MAD algorithm is twofold. Firstly, it targets noise reduction to facilitate the human monitoring/inspection procedure. Secondly, it aims to improve the efficiency/accuracy of the consecutive processing steps, namely, video compression and motion detection. Motion-based video surveillance techniques were modified to the specific needs of human activity monitoring. As a result, the "wavelet-domain dynamic background/foreground segmentation" procedure was developed in combination with the "wavelet-domain empirical Wiener filtering" video denoising technique. The computational efficiency of the proposed work relies on the fact that a single methodology accomplishes the two different tasks: video enhancement and motion detection, with the advantage of reduced computational load when compared to motion-vector-based motion estimation approaches. The method was tested in a video-assisted biomedical monitoring application and it was proved to efficiently work under poor lighting conditions and significant noise problems. Based on the qualitative and quantitative analysis results, the proposed methodology is expected to be easily extendible to similar video surveillance tasks, as well as in demanding denoising and multimedia content management applications. Future work involves extension and full automation of the dynamic BRFR reinitiation process, improvements towards more efficient video denoising and development of video compression algorithms. Video denoising comparisons of the proposed methodology with classical algorithms using standard testing sequences are in preparation for publication. In the semantic characterization domain, an MPEG-7 schema for the accommodation of biomedical-assisting audiovisual content is currently under development. Further implementation includes extensions to psychophysiological monitoring areas (i.e., task-performance analysis) and general human activity applications.

ACKNOWLEDGMENT

The authors wish to thank Dr. A. Kalampakas for his valuable contribution during the experimental phase of the work.

REFERENCES

[1] C. A. Dimoulas, "Audio-visual processing and content management techniques, for the study of (human) bioacoustics' phenomena," Ph. D. dissertation, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece, November 2006.

[2] M. J. Davey, "Investigation of sleep disorders," *Journal of Paediatrics and Child Health*, vol. 41, no. 1-2, pp. 16–20, 2005.

[3] J. C. T. Pepperell, R. J. O. Davies, and J. R. Stradling, "Sleep studies for sleep apnoea," *Physiological Measurement*, vol. 23, no. 2, pp. R39–R74, 2002.

[4] Z. Li, A. M. da Silva, and J. P. S. Cunha, "Movement quantification in epileptic seizures: a new approach to video-EEG analysis," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 6, pp. 565–573, 2002.

[5] M. A. Coyle, D. B. Keenan, L. S. Henderson, et al., "Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease," *Cough*, vol. 1, no. 3, pp. 1–7, 2005.

[6] M. J. Hensley, D. R. Hillman, R. D. McEvoy, et al., "Guidelines for sleep studies in adults," in *The Australasian Sleep Association & Thoracic Society of Australia and New Zealand*, pp. 1–38, Sydney, Australia, October 2005.

[7] K. Nakajima, Y. Matsumoto, and T. Tamura, "Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed," *Physiological Measurement*, vol. 22, no. 3, pp. N21–N28, 2001.

[8] T. Josefsson, E. Nordh, and P.-O. Eriksson, "A flexible high-precision video system for digital recording of motor acts through lightweight reflex markers," *Computer Methods and Programs in Biomedicine*, vol. 49, no. 2, pp. 119–184, 1996.

[9] J. C. Guerri, M. Esteve, C. Palau, M. Monfort, and M. A. Sarti, "A software tool to acquire, synchronise and playback multimedia data: an application in kinesiology," *Computer*

Methods and Programs in Biomedicine, vol. 62, no. 1, pp. 51–58, 2000.

[10] S. Zeng, J. R. Powers, and H. Hsiao, "A new video-synchronized multichannel biomedical data acquisition system," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 3, pp. 412–419, 2000.

[11] N. B. Karayiannis and G. Tao, "An improved procedure for the extraction of temporal motion strength signals from video recordings of neonatal seizures," *Image and Vision Computing*, vol. 24, no. 1, pp. 27–40, 2006.

[12] C. A. Dimoulas, G. M. Kalliris, G. V. Papanikolaou, and A. Kalampakas, "Long-term signal detection, segmentation and summarization using wavelets and fractal dimension: a bioacoustics application in gastrointestinal-motility monitoring," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 438–462, 2007.

[13] C. A. Dimoulas, G. M. Kalliris, G. V. Papanikolaou, V. Petridis, and A. Kalampakas, "Bowel-sound pattern analysis using wavelets and neural networks with application to long-term, unsupervised, gastrointestinal motility monitoring," *Expert Systems with Applications*, vol. 34, no. 1, pp. 26–41, 2008.

[14] R. L. Lagendijk, P. M. B. van Roosmalen, and J. Biemond, "Video enhancement and restoration," in *Handbook of Image and Video Processing*, J. D. Gibson and A. C. Bovik, Eds., pp. 227–241, Academic Press, San Diego, Calif, USA, 2000.

[15] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 113–118, 2005.

[16] F. Jin, P. Fieguth, and L. Winger, "Wavelet video denoising with regularized multiresolution motion estimation," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 72705, 11 pages, 2006.

[17] V. Zlokolica, A. Pizurica, and W. Philips, "Wavelet-domain video denoising based on reliability measures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 8, pp. 993–1007, 2006.

[18] E. J. Balster, Y. F. Zheng, and R. L. Ewing, "Combined spatial and temporal domain wavelet shrinkage algorithm for video denoising," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 220–230, 2006.

[19] F. Pereira and P. Salembier, Eds., "Special issue on MPEG-7," *Signal Processing: Image Communication*, vol. 16, no. 1-2, pp. 1–293, 2000.

[20] "Special issue on MPEG-7," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 11, no. 6, pp. 685–772, 2001.

[21] P. Salembier, "Overview of the MPEG-7 standard and of future challenges for visual information analysis," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 4, pp. 343–353, 2002.

[22] J. Calic and E. Izquierdo, "Temporal segmentation of MPEG video streams," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 6, pp. 561–565, 2002.

[23] I. Yahiaoui, B. Merlaldo, and B. Huet, "Comparison of multiepisode video summarization algorithms," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 48–55, 2003.

[24] A. Divakaran, R. Radhakrishnan, and K. A. Peker, "Video summarization using descriptors of motion activity: a motion activity based approach to key-frame extraction from video shots," *Journal of Electronic Imaging*, vol. 10, no. 4, pp. 909–916, 2001.

[25] R. T. Collins, A. J. Lipton, T. Kanade, et al., "A system for video surveillance and monitoring: VSAM final report," Tech. Rep. CMURI-R-00-12, Carnegie Mellon University, Pittsburgh, Pa, USA, 2000.

[26] B. U. Töreyin, A. E. Çetin, A. Aksay, and M. B. Akhan, "Moving object detection in wavelet compressed video," *Signal Processing: Image Communication*, vol. 20, no. 3, pp. 255–264, 2005.

[27] J. Konrad, "Motion detection and estimation," in *Handbook of Image and Video Processing*, J. D. Gibson and A. C. Bovik, Eds., pp. 207–225, Academic Press, San Diego, Calif, USA, 2000.

[28] D. E. Butler, V. M. Bove Jr., and S. Sridharan, "Real-time adaptive foreground/background segmentation," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 14, pp. 2292–2304, 2005.

[29] B. Erol and F. Kossentini, "Retrieval by local motion," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 41–47, 2003.

[30] C. A. Dimoulas, G. M. Kalliris, G. V. Papanikolaou, and A. Kalampakas, "Novel wavelet domain wiener filtering denoising techniques: application to bowel sounds captured by means of abdominal surface vibrations," *Biomedical Signal Processing and Control*, vol. 1, no. 3, pp. 177–218, 2006.

[31] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, Calif, USA, 2nd edition, 1999.

[32] D. Dong and A. C. Bovik, "Wavelet denoising for image enhancement," in *Handbook of Image and Video Processing*, J. D. Gibson and A. C. Bovik, Eds., pp. 117–123, Academic Press, San Diego, Calif, USA, 2000.

[33] A. De Stefano, P. R. White, and W. B. Collis, "Training methods for image noise level estimation on wavelet components," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 16, pp. 2400–2407, 2004.

[34] E. J. Balster, Y. F. Zheng, and R. L. Ewing, "Feature-based wavelet shrinkage algorithm for image denoising," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2024–2039, 2005.

[35] M. A. Santiago, G. Cisneros, and E. Bernues, "Iterative desensitisation of image restoration filters under wrong PSF and noise estimates," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 72658, 18 pages, 2007.

[36] V. Bruni and D. Vitulano, "Old movies noise reduction via wavelets and wiener filters," *Journal of WSCG*, vol. 12, no. 1–3, pp. 8 pages, 2004.

[37] A. Pizurica, V. Zlokolica, and W. Philips, "Combined wavelet domain and temporal video denoising," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '03)*, pp. 334–341, Miami, Fla, USA, July 2003.

[38] C. A. Dimoulas, C. Vegiris, K. A. Avdelidis, G. M. Kalliris, and G. V. Papanikolaou, "Automated audio detection, segmentation, and indexing with application to postproduction editing," in *Proceedings of the 122nd Audio Engineering Society Convention*, no. 7138, Vienna, Austria, May 2007.

[39] C. A. Dimoulas, G. M. Kalliris, C. Sevastiadis, G. V. Papanikolaou, and D. Christidis, "Development of an engineering application for subjective evaluation of human response to noise," in *Proceedings of the 110th Audio Engineering Society Convention*, no. 5408, Amsterdam, The Netherlands, May 2001.

[40] J. M. Ferryman, "Performance metrics and methods for tracking in surveillance," in *Proceedings of the 3rd IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS '02)*, E. Tim, Ed., pp. 26–31, Copenhagen, Denmark, June 2002.