A Three-Dimensional Spatiotemporal Template for Interactive Human Motion Analysis

Alexandra Branzan Albu
Department of Electrical and Computer Engineering, Victoria (BC), Canada
Email: aalbu@ece.uvic.ca

Trevor Beugeling
Department of Electrical and Computer Engineering, Victoria (BC), Canada
Email: trjb@uvic.ca

***Abstract*—This paper describes a new three-dimensional spatiotemporal template, namely the Volumetric Motion History Image (VMHI), for the purpose of human motion analysis. Irregularities in human actions typically occur either in speed or orientation; they carry information about the balance and the confidence level of the human subject performing the activity. The proposed VMHI template handles successfully shortcomings of existing spatiotemporal templates related to motion self-occlusion and speed. Therefore, VMHI allows for interactive visualization, as well as quantification of motion performance. This study focuses on the analysis of sway and speed-related abnormalities, which are among the most common motion irregularities in the studied set of human actions.**

***Index Terms*—computer vision, interactive motion analysis, spatiotemporal motion representation**

I. INTRODUCTION

Human motion analysis has been a central research topic in the Computer Vision community for over two decades. As shown in a recent survey by Moeslund et al [1], this sustained interest is motivated by the theoretical complexity of the problem (i.e. the high variability and level of sophistication of human gestures and activity), and also by the wide spectrum of potential applications in biometrics, medicine, sports, gesture-controlled interfaces etc.

The analysis of human motion from video data is mostly focused by applications in *surveillance*. Indeed, prestigious academic journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and the Journal on Multimedia Systems have dedicated special issues to research advances in vision-based surveillance [2], [3]. Some major surveillance-related themes address human activity identification, gait-based biometrics and real-time abnormal event detection.

An emerging area of interest for vision-based human motion analysis is the field of *perceptual human-computer interfaces*. Perceptual interfaces are typically multi-modal; their design is focused towards broadening the bandwidth of communication between the user and the system. Computer Vision techniques for gesture recognition have been successfully integrated into perceptual interfaces for video games such as EyeToy [4], and for motor-impaired users, as proposed by Betke et al in [5].

Both above-mentioned areas of application, namely video surveillance and perceptual human-computer interaction, are detection- and recognition-oriented. The main goal in such applications is to detect and recognize either the motion/action/gesture performed by the human, or the human subject herself from her motion-based signature.

However, changing the main goal from recognition to analysis and quantification of motion performance unveils a variety of new challenges and new research directions for the development of computer vision algorithms. In other words, the main question that we want to investigate is "How well is the motion performed?" rather than "What motion is performed?". At the best of our knowledge, motion analysis for performance quantification is still a fairly unexplored field in computer vision with promising application areas such as aging-in-place, rehabilitation, sports etc.

The work described in this paper was done in a rehabilitation context about frail elderly subjects. Our main goal is not to recognize the activity performed by a subject, but to analyze a standardized set of common human activities in order to quantify and monitor the subjects' performance over time. In our research, abnormal motion is not considered an outlier, but a quantifiable deviation from normal motion.

Motion analysis is not a well-studied problem; indeed, it poses different challenges with respect to action or subject recognition. For instance, early work of Cutting and Kozlowski [6] showed that typical humans are experts in recognizing human activities and perform well at identifying familiar people from their gait. Therefore,

Based on "Analysis of Irregularities in Human Actions Using Volumetric Motion History Images", by A. Branzan Albu, T. Beugeling, N. Virji-Babul, and C. Beach which appeared in the Proceedings of the IEEE International Workshop on Motion and Video Computing, Austin (TX), US, February 2007. © 2007 IEEE.

vision-based algorithms for activity recognition and gait identification can easily be validated against ground truth data produced by human reasoning. In fact, intelligent visual surveillance systems aim to be a more efficient alternative to CCTV systems with a human operator in the loop.

On the other side, an accurate performance analysis of basic, daily human activities can be performed only by professionals such as physiotherapists and kinesiologists. For this purpose, they typically employ measurements such as the Berg Balance Score (BBS) [7] which assesses on a 5 point ordinal scale 14 basic human activities common to everyday life. Hence, the validation of a computer vision-based approach for human motion assessment against BBS is not feasible. This new type of approach should be validated with physiotherapists in the loop and designed as to allow an optimal trade-off between interactivity and automation. It is anticipated that an interactive tool based on computer vision algorithms for human motion analysis will complement score-based assessments as it will provide additional information about the subject's evolution over time.

This paper proposes a new spatiotemporal template, namely the Volumetric Motion History Image (VMHI), and its application to the analysis of irregularities in human motion. Preliminary results of our study appeared in Branzan Albu et al [20]. The present paper contains significant conceptual and experimental updates with respect to [20].

The remainder of the paper is organized as follows. Section 2 presents related work in the field of human motion analysis. Section 3 describes our proposed approach, while section 4 discusses experimental results. Section 5 draws conclusions and outlines ongoing and future work directions.

II. RELATED WORK

The extraction of information about human motion from video sequences can be performed with a diversity of approaches, which can be either model-based or model-free, as shown in Aggarwal and Cai [8]. This motion information is necessary for building a motion representation appropriate for the task of interest. While motion representation is identified by Moeslund and Granum [9] as an essential component of tracking (a pre-processing step), it is also necessary for the achievement of global goals such as recognition and analysis. The generation of an adequate motion representation is not trivial, since trade-offs must be made between the richness of the representation, the time necessary for generating it and the computational complexity of the algorithms which will further use it. These trade-offs are usually critical for real-time decision systems. Other challenges are related to the robustness of the motion representation to self-occlusion and to errors occurred in background subtraction, tracking, or other preprocessing steps. As a conclusion, the motion representation must be task-oriented in order to address properly all contextual constraints.

The rehabilitation context of our work imposes constraints such as dealing with loose clothing and with a high variability of body shapes. For this reason, model-based approaches for motion representation do not represent a suitable option, since they are based on joint segmentation and tracking. The remainder of this section will therefore discuss related work on holistic or appearance-based motion representations.

Appearance-based methods focus on whole-body motion of the human silhouette, without decomposing it into absolute and the relative motion of body parts. These methods identify global patterns of motion as opposed to temporal trajectories of anatomic joint; these patterns are usually encoded as spatiotemporal templates.

Most of existing spatiotemporal templates are 2D images, which offer a compact representation of motion information and are suitable for analysis and classification using standard image processing techniques for feature extraction and pattern recognition.

Polana and Nelson proposed in [10] the temporal texture for the study of quasi-random motion such as windblown trees and ripples on water; the temporal texture can be further analyzed with standard techniques similar for spatial texture analysis. They also introduced in [11] a feature-based representation of periodic human motion; this representation contains spatiotemporal motion magnitudes obtained with Fourier image analysis. Cutler and Davis introduced in [12] the 2D inter-frame similarity matrix, a spatiotemporal representation that they used for detecting periodic human and non-human motion.

While periodicity characterizes all types of human locomotion (walking and running), the above-mentioned approaches for motion representation are not extendable for the analysis of non-periodic basic human actions. The concepts of Motion Energy Image (MEI) and Motion History Image (MHI) introduced by Davis and Bobick in [13] are spatiotemporal templates useful for aperiodic human activity description and recognition. MEI is a binary image which encodes the spatial occurrence of motion throughout the video sequence. MHI is a gray-level image which encodes the recency of motion in gray-levels. Both MEI and MHI are generated over a temporal window of τ frames, with parameter τ either empirically chosen or computed after an exhaustive, iterative matching against a reference template. Moreover, MEI and MHI are not independent, since MEI can be retrieved from MHI via a simple thresholding.

The main advantage of the MHI representation is its compactness, which makes it suitable for real-time activity recognition. However, the recognition process has to be thoroughly supervised, since it needs to be label-based and reference templates must be available for each activity of interest. A second shortcoming of the initial MHI representation is its lack of robustness against spatial motion self-occlusion occurring during the same temporal window; this event happens rather frequently in human actions.

A hierarchical extension to the original MHI framework was proposed by Davis in [14]. This extension

aims at eliminating previous problems related to limited recognition capabilities and variable speed of motion. A second parameter δ is introduced in [14] for measuring the decay factor, which is necessary for varying the length of the captured history of movement. The hierarchical pyramid of MHIs allows for recovering to a certain extent motions of varying speed by exploiting spatial gradient information.

Valstar et al described in [15] a different extension of MHI, namely the multiple-level MHI (MMHI), which aims at handling motion self-occlusion by recording motion history at multiple time intervals. Their work focused on the automatic detection of facial actions units that compose facial expressions. The experimental results shown in [15] do not clearly demonstrate the superior performance of MMHI with respect to the standard MHI in the context of their application.

Weinland et al introduced in [16] a 3D extension to the initial MHI, namely Motion History Volumes. This extension was used for viewpoint-independent action recognition. The transition from 2D to 3D is straightforward, since pixels are replaced with voxels, and the standard image differencing function D(x,y) is substituted with an occupancy function D(x,y,z,t). The Motion History Volumes offer an interesting alternative to action recognition from video stream acquired simultaneously with multiple cameras. However, issues such as the additional computational complexity introduced by calibration, synchronization of multiple cameras, and parallel background subtraction are not discussed in [16].

Yilmaz and Shah [17] use spatiotemporal volumes for action recognition. The volumes are created by stacking silhouette contours extracted from adjacent frames into parallel equidistant planes. Inter-slice point-by-point correspondences are obtained using weighted bipartite graphs. Their motion representation, which is a parametric 3D surface, is described using 8 surface primitives, namely peaks, ridges, saddle ridges, pits, valleys, and saddle valleys. The types and relative locations of the surface primitives on the Spatiotemporal Volume corresponding to one atomic action compose an action sketch.

Blank et al [18] also use spatiotemporal volumes for the description and recognition of human actions. They build the volumes by assigning to each space-time point the mean time required for a particle undergoing a random-walk process starting from the point to hit the boundaries. Their action representation is based on primitive spatio-temporal entities, namely "sticks", "plates", and "balls". Spatio-temporal saliency is also extracted at every point in the shape, which allows for minimizing the number of features representing an action.

When shifting the primary focus from motion recognition to motion analysis, existing 2D and 3D motion representations exhibit shortcomings. For instance, it has been shown in [17, 18] that the action sketch and the space-time shapes are able to discriminate between different actions such as dancing, kicking, walking, jumping-jacks etc. However, the current form of the above representations does not allow for detecting and quantifying subtle differences that occur between various performances of the same action.

To address such current limitations, this paper describes a new methodology for the interactive analysis of a volumetric motion representation. Our motion representation, called Volumetric Motion History Image (VMHI), is a 3D extension for the initial MHI concept in [13]. Our proposed approach allows for interactive visualization and analysis. Moreover, it handles successfully issues such as motion self-occlusion, speed variability, and variable-length motion sequences. The following section describes in detail the proposed motion representation as well as a new measure for motion irregularity.

III. Proposed Approach

A. The Volumetric Motion History Image

The standard MHI concept is computed in [13] using an iterative replacement and decay operator as follows:

$$H_\tau(x,y,t)=\begin{cases}\tau \ \ if \ \ D(x,y,t)=1\\ \max(0,H_\tau(x,y,t-1)-1) \text{ otherwise}\end{cases} \quad (1)$$

$$MHI(x,y)=H_\tau(x,y,\tau)$$

where:

- D(x,y,:) is a binary image sequence indicating regions of motion and created with a simple frame-by-frame differencing technique.
- τ is the length of the temporal window.

The second equation in (1) does not appear explicitly in [13]; we have derived it after a close analysis of their experimental results and of their MHI example illustrations.

Our paper proposes a 3D extension of MHI which eliminates the need of a pre-specified length of the temporal window τ. Input data is represented by an image sequence S(:, :, k) k=1..N of binary silhouettes obtained with background subtraction from an initial video sequence acquired with one stationary camera.

The proposed Volumetric Motion History Image (VMHI) is a set of parallel and equidistant slices where the z coordinate encodes discrete temporal information as represented by frame indexes. In this sense, it is similar to the spatiotemporal volume used in [17], although it neither stacks silhouette contours, nor computes inter-slice point-by-point correspondences.

Let us consider *contS* (:, :, *k*), the one pixel thick contour of the binary silhouette in frame *k*. The VMHI representation is defined as follows:

$$VMHI(x,y,k)=\begin{cases}S(x,y,k)\Delta S(x,y,k+1)\\ if \ contS(x,y,k)\neq contS(x,y,k+1)\\ 1 \ if \ contS(x,y,k)=contS(x,y,k+1)\end{cases} \quad (2)$$

$$x=\overline{1,X}, y=\overline{1,Y}, k=\overline{1,N-1}$$

where *x, y* correspond to spatial coordinates in the image plane and X,Y are the frame dimensions in pixels, while k encodes discrete temporal information; Δ stands for the symmetric difference operator.

Each slice in the VMHI representation is built by integrating two types of information, related to:

a) the motion occurred within a pair of adjacent frames, captured with the symmetric difference operator between two adjacent binary silhouettes;

b) the spatial occupancy, captured with the binary contour comparison.

The standard MHI [13] is based on motion information only, which is retrieved by using a simple frame-by-frame differencing technique. Its hierarchical extension in [14] can be computed using either motion information or spatial occupancy; the spatiotemporal volumes in [17, 18] are based on spatial occupancy only. We believe that integrating information about both motion and spatial occupancy can provide a more robust representation than using one source of information only. The use of spatial occupancy only results in connected spatial regions in the VMHI horizontal slices; however, it does not lead to an explicit motion representation. The use of silhouette differencing only for the extraction of motion information leads to disconnected regions in the horizontal slices of the VMHI, which are difficult if not impossible to visualize in 3D.

Fig. 1 shows an example of articulated human motion. The binary image in Fig. 1b contains motion information only obtained by upper body silhouette differencing. Due to motion self-occlusion and imperfect background subtraction, a significant portion of the arm contour is lost. These portions are successfully retrieved in Fig. 1c, which shows a slice of VMHI computed with (1). The gray level display is used for the purpose of visualization only: it shows contour information (spatial occupancy) in white, motion occurring from background-foreground transition in light gray, and motion from foreground-background transition in dark gray.

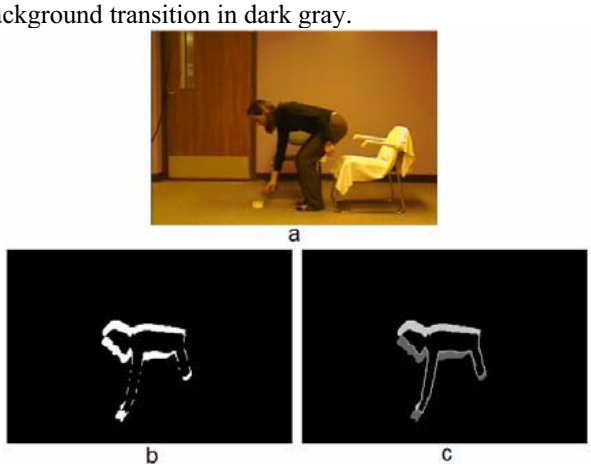

Fig. 1. a) frame in a sequence containing a picking up action; b) result of binary silhouette differencing; c) result obtained with Eq.(1).

One may notice the disappearance of the τ parameter encoding the length of the temporal window from the VMHI definition. In the standard (MEI, MHI) framework for activity recognition, τ was either empirically set, or determined with an exhaustive research of the best correlation match between the template to be classified and a reference template. The absence of a reference template and the change of goal from recognition to analysis led to the conclusion that τ is not needed in the VMHI representation. The temporal window of interest is typically defined over the entire motion sequence; it can also be specified by the user via a graphical user interface.

B. Overcoming limits of the MHI representation in the motion analysis context

The (MEI, MHI) template set was proven reliable for action recognition in controlled environments such as KidsRoom[19]. However, this compact motion representation is not able to capture subtle details of human motion required for an accurate quantitative and qualitative motion analysis. The main factors limiting the use of (MEI, MHI) representation in a motion analysis context are listed below. The ability of the proposed VMHI to overcome these shortcomings is also discussed.

1) Motion self-occlusion.

Due to the replacement-and decay operator used in (1) for the computation of the standard MHI [13], the most recent motion will overwrite all the motion information previously gathered at the same spatial location. In the context of motion analysis, this overwriting process leads to loss of important information and has to be eliminated.

To analyze systematically the effects of the overwriting process on the MHI, we have built two sequences of identical length (58 frames) containing rigid horizontal translation with different motion irregularities. Key frames of the two sequences are shown in Fig. 2a. Sequence A contains a rectangle in translation; its motion changes orientation for a certain time interval (frames 11-31), then resumes the initial orientation. Sequence B is identical to A, except for the irregularity occurring during the same time interval (frames11-31), where the object stops and remains immobile instead of moving leftwise. The computation of the MHI was performed as in [13] with τ=58. Since both sequences have identical MHIs, it can be concluded that the MHI representation fails to capture information about irregularities in motion. The choice of a smaller τ can certainly lead to different MHIs for the two test sequences, but this choice would have to be empirically made, since no *a priori* information about the occurrence and nature of irregularities is available.

Fig. 2c and 2d contain the VMHI motion representations for test sequences A and B respectively. A simple visualization of these 3D motion models allows for: a) detecting the occurrence of the motion irregularity in both sequences; b) discriminating between the two motion irregularities.

One may notice that capturing a temporary stop of the moving object (as in test sequence B) is possible because the VMHI model definition in Eq. (2) integrates information about motion and spatial occupancy. Sudden short stops occur quite frequently in actions performed by elderly subjects, as they usually correspond to hesitations of the subject.

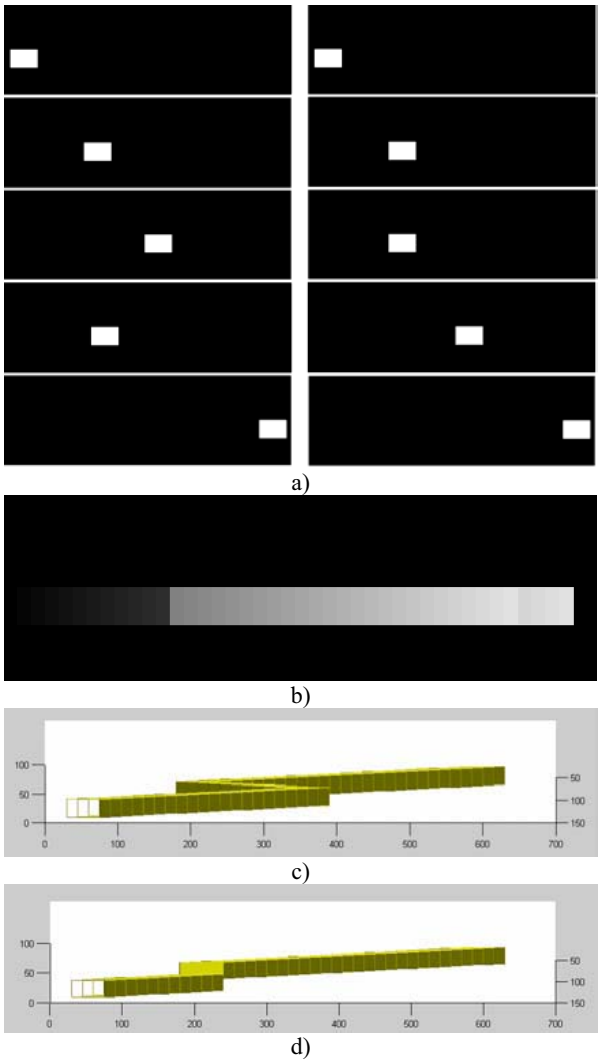

Figure 2. a) left: keyframes corresponding to the test sequence A containing a temporary change in motion orientation; right: keyframes corresponding to the test sequence B containing a temporary stop of the moving object; b) MHI is identical for both test sequences; c) VMHI for sequence A; d) VMHI for sequence B.

2) Subject-dependent motion speed.

While fit subjects perform a given activity in a relatively short time frame, hesitant frail subjects take usually a much longer time to perform the same activity. For instance, at an acquisition rate of 30 fps, the length of a typical sit-to-stand normal action sequence is 56 frames, while an abnormal sit-to-stand lasts 148 frames. Thus, it is difficult to quantify the differences between normal and abnormal actions using the standard MHI, since this representation is defined over a fixed-length temporal window. An attempt to normalize the length of the sequences before the MHI generation has been reported in [15] for facial action units. The proposed normalization was based upon a uniform subsampling of the longer sequence. The spatiotemporal volumes in [17] are also speed-normalized, since Yilmaz and Shah claim that volumes obtained from an initial action sequence and from a randomly subsampled version of the same action look similar. However, neither uniform nor random undersampling are appropriate for the context of our work, since by eliminating frames in the longer sequence relevant information about motion irregularities will be lost.

Our proposed solution is the computation of VMHI models for normal and abnormal motions respectively without attempting to normalize the models for a direct inter-model comparison. The irregularities the spatiotemporal 3D surface of each model are to be interactively observed and selected via a graphical user interface and further analyzed with a measure of surface smoothness (see Section IV). The variable speed of motion, which is also an index of abnormality, can also be directly measured if working with non-normalized models. Such measurements are detailed in Section IV.

Fig. 3 shows VMHI models for normal and abnormal sit-to-stand. Key frames of the abnormal and normal motion are shown in Fig. 3a and 3b respectively. For the abnormal motion, the key frames capture a significant motion irregularity occurring at the beginning of the action, namely a horizontal sway used by a frail subject to initiate the upward motion. The VMHI model in Fig. 3d corresponds to the temporal interval where this irregularity occurs. The normal motion does not present irregularities and therefore results in a smooth VMHI.

3) Variable length of the action sequences.

The capture of slow motion results in long video sequences. Therefore, an increase in the length τ of the temporal window is translated into a larger number of gray levels in the MHI. Encoding motion information in a gray-level image is thus limited by the maximum number of gray levels. The proposed VMHI representation encodes the temporal information along the z coordinate, which is a suitable solution for both long and short sequences.

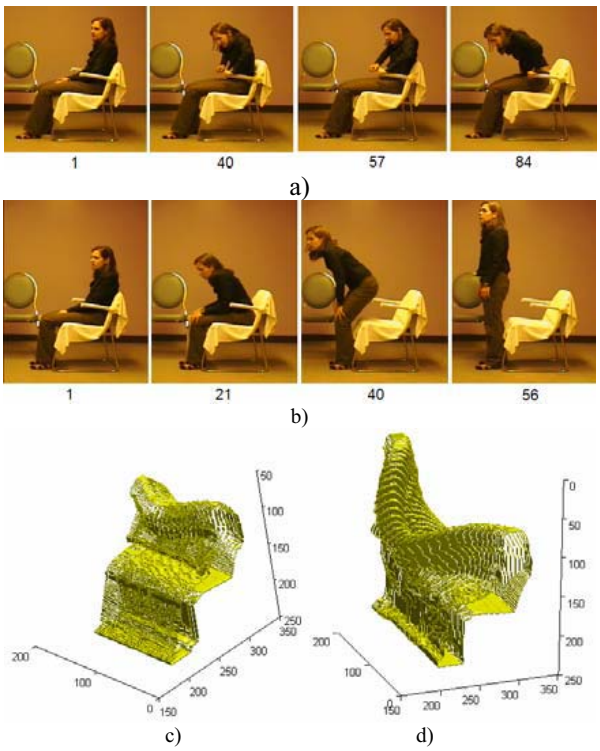

Figure 3. a) key frames in abnormal sit-to-stand showing initial sway; b) key frames in normal sit-to-stand; c) VMHI for the initial sway in abnormal sit to stand; d) VMHI for the normal sit-to-stand sequence.

4) Motion periodicity

Due to the overwriting phenomenon discussed above, the MHI template is not applicable for the analysis of periodic human motion. The approaches in [17, 18] are also applicable to atomic, non-periodic actions only, due to the automatic extraction of surface-related descriptors.

The proposed VMHI enables the representation and analysis of periodic motion. Moreover, it can be also used for detecting the fundamental period of motion by searching for pairs of similar slice sequences in the VMHI model separated by a minimum number of frames.

IV. Experimental Results

A. Design of the experiment

The input data for the proposed work consists in 6 pairs of video sequences containing normal and abnormal motion respectively. Each pair comprises the same human action. The actions of interest, selected in collaboration with a team of domain experts (physiotherapists) consist of sit-to-stand, stand-to-sit, reaching, picking up an object placed on the floor, stepping on one stair step, and transfer from one chair to another. Both normal motions and abnormal simulations were performed by a certified physiotherapist. Working with simulated abnormal actions was considered most appropriate for the goal of this study which focused on finding the optimal motion representation for the analysis of abnormal motion. As mentioned in Section V, future work will focus on the analysis of the motion of frail elderly subjects. All sequences in our database were acquired with a monocular camera at 30 frames/second from an orthogonal view to the direction of motion. Prior to generating VMHI models, all sequences were binarized with a background subtraction method based on statistical foreground-background differences.

In the context of the work presented here, abnormal motion is defined as a quantifiable deviation from normal motion. This deviation is due to several types of spatiotemporal motion irregularities. The sway is the most encountered motion irregularity and consists in repetitive, quick changes in motion orientation due to a temporary loss of balance or to insufficient strength in lower limbs (see Fig. 3a for an example). Other motion irregularities include variations in the speed magnitude, temporary stops in motion, and limited range of motion.

While the VMHI representation allows for the visualization and quantification of all the above mentioned motion irregularities, the work presented in this paper is focused on sway and speed analysis only.

B. Sway analysis and visualization

Human motion can be defined from a kinematic standpoint as an articulated motion, as it is composed of constrained relative translations/rotations of the various body parts. A smooth, consistent limb/torso translation or rotation defines a quasi-planar surface region in its VMHI template; the orientation of this surface region encodes the direction of motion. Therefore, the VMHI surface of a normal, temporally smooth motion is spatially smooth, and piecewise planar. It consists of a limited number of quasi-planar surface regions with orientations encoding the direction of motion of various body parts. The unit normals to the vertices in each quasiplanar region exhibit therefore a low variance in their orientation.

In a motion with sway-type irregularities, the relative translation of body parts features frequent changes of speed and orientation. Consequently, its corresponding VMHI features an 'unsmooth' appearance, since the sway translates into spatiotemporal 'ripples'. The normals to the vertices in the VMHI surface region corresponding to a sway have therefore a high variance in their orientation.

The above observations resulted in choosing a descriptor of the VMHI surface smoothness for the analysis of sway irregularities. For a given activity performed abnormally (i.e. as *abnormal activity*), this descriptor measures its deviation from the same activity performed normally (i.e. the *normal activity*).

Let us consider the following notations:

- $VMHI_{abn}$: the VMHI of the abnormal activity;
- $VMHI_n$: the VMHI of the normal activity;
- (n_x, n_y, n_z): unit normal vectors defined on each vertex of the VMHI surface.
- $\text{var}(n_x, n_y, n_z)|_{VMHI} = (\text{var}(n_x), \text{var}(n_y), \text{var}(n_z))|_{VMHI}$: the statistical variance of the orientations of the normal vectors to a given VMHI surface.

The deviation of an abnormal activity from its corresponding normal activity is measured with a deviation vector D, defined as follows:

$$D = \lfloor D_x, D_y \rfloor$$

$$D_x = \frac{\text{var}(n_x)\big|_{VMHI_{abn}} - \text{var}(n_x)\big|_{VMHI_n}}{\text{var}(n_x)\big|_{VMHI_n}} \qquad (3)$$

$$D_y = \frac{\text{var}(n_y)\big|_{VMHI_{abn}} - \text{var}(n_y)\big|_{VMHI_n}}{\text{var}(n_y)\big|_{VMHI_n}}$$

The D_x and D_y components of the deviation vector quantify motion irregularities occurring along the vertical and horizontal directions respectively (i.e. horizontal and vertical sway).

The deviation vector has only two components, since the variance of n_z, var (n_z) is significantly influenced by the speed of motion. The following experiment is to prove this affirmation. Three test sequences containing a 60x40 pixel rectangle in horizontal translation at constant slow, medium and fast speed were generated. As shown in Table 1, the variance var (n_z) of the unit normal vectors over the VMHI templates generated for each sequence is clearly correlated with the speed of motion.

Table 1. Correlation between var (n_z) and speed of translatory motion in three test sequences.

Speed of motion	var(n_z)
2 pixels/frame	0.3775
5 pixels/frame	0.6217
10 pixels/frame	0.7684

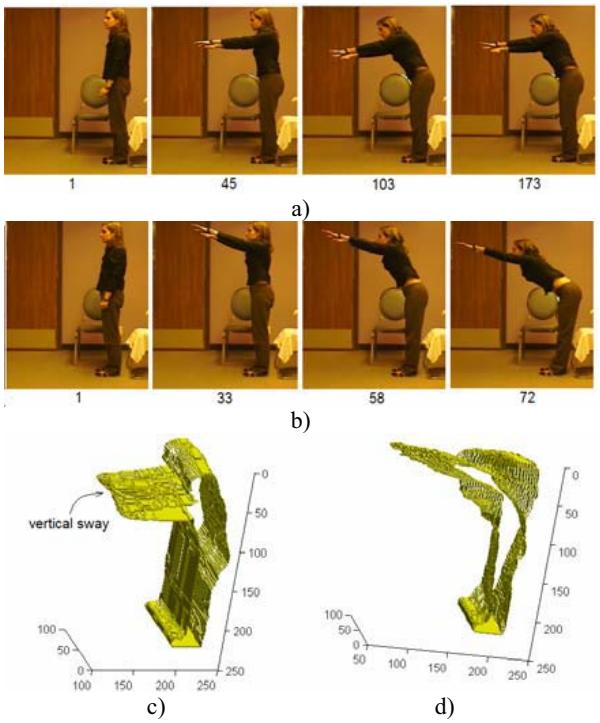

Figure 4. a) key frames in abnormal reaching; b) key frames in normal reaching; c) VMHI for user-selected sway region in abnormal reaching ; d) VMHI for normal reaching.

As mentioned in section 3.2, this work does not deal with the automatic detection of sway-type irregularities. Instead, the VMHI model is adequate for interactive visualization, modification and analysis of sway in a user interface which enables the user to:

a) rotate freely the VMHI of a given human action for inspection of motion irregularities;

b) visualize simultaneously and thus compare two VMHIs of normal and abnormal motion respectively;

c) generate a more precise VMHI for a spatiotemporal region of interest containing sway by specifying its temporal limits (i.e. start and end frames);

d) compute the deviation vector D between normal and abnormal motion with Eq. (2)

Figures 3, 4, 5, 6, and 7 illustrate the visualization of VMHI templates corresponding to normal and abnormal actions respectively. Sway-type motion irregularities are annotated in each figure.

Table 2 contains the statistical variance of the normal orientations to the surfaces of the VMHI models built for each activity. It is easy to notice that var(n_x) and var(n_y) are always higher in the abnormal activity than in the corresponding normal activity. This result is consistent with our initial assumption which correlates the smoothness of the VMHI surface to the motion coherency.

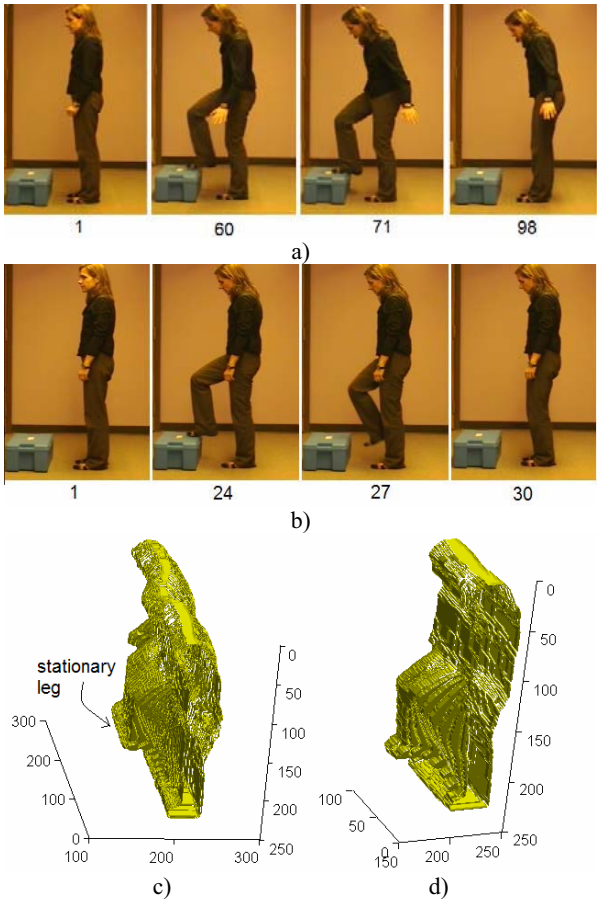

Figure 5. a) key frames in abnormal stepping; b) key frames in normal stepping; c) VMHI for user-selected region in abnormal stepping showing a temporary stop in motion; d) VMHI for normal stepping.

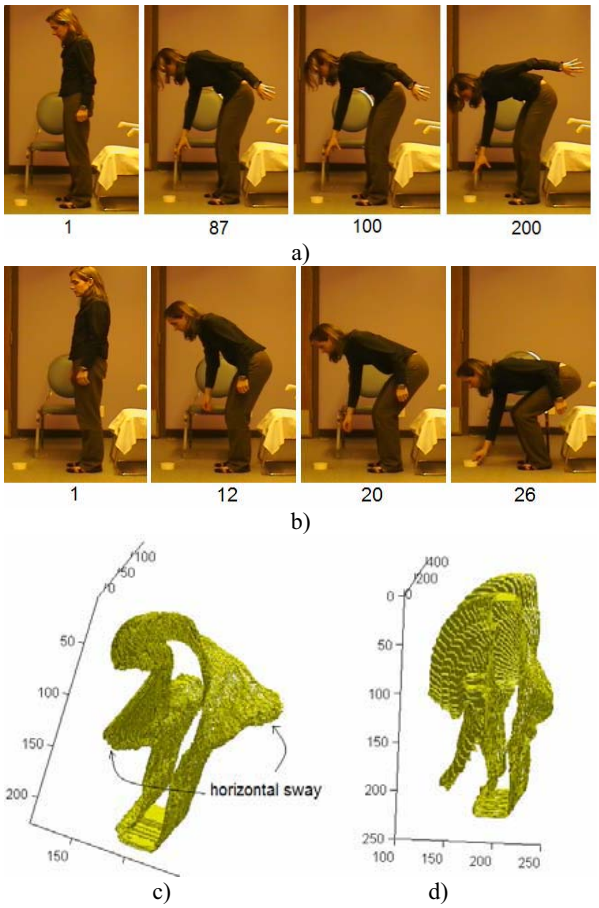

Figure 6. a) key frames in abnormal picking up; b) key frames in normal picking up; c) VMHI for user-selected region in abnormal picking up showing horizontal sway; d) VMHI model for normal picking up.

Table 3 uses the data in Table 2 for computing the $[D_x, D_y]$ vector of deviations with Eq. (2). The largest deviation, D_y=76.41%, was obtained for the horizontal sway in the abnormal sit-to-stand. This result is consistent with the qualitative observation of a large-amplitude sway in the abnormal sit-to-stand video. The abnormal reaching, characterized by a subtler vertical sway, has the D_x deviation larger than D_y. The abnormal picking up and the stand-to-sit feature horizontal sway, which results in D_y larger than D_x. The smallest D_x and D_y deviations are obtained for the abnormal stepping, which contains a temporary stop.

The abnormality in the transfer motion (see Fig. 7) is not represented by sway; the temporal evolution of postures is what differentiates abnormal from normal in this case. Indeed, in abnormal transfer (see Fig. 7a) the subject assumes a much lower posture than in normal transfer (see Fig. 7b). The postural differences are visible in Fig. 7c and 7d. One may also notice that VMHI is suitable for analyzing human actions which involve rotations around the body axis, which is not feasible with other motion representations built from monocular sequences.

The above results suggest that, for temporary stops in motion and for motions differentiated by postural information, additional measures, for instance speed-related, are necessary. The next section describes a simple measure for the speed of motion using VMHI.

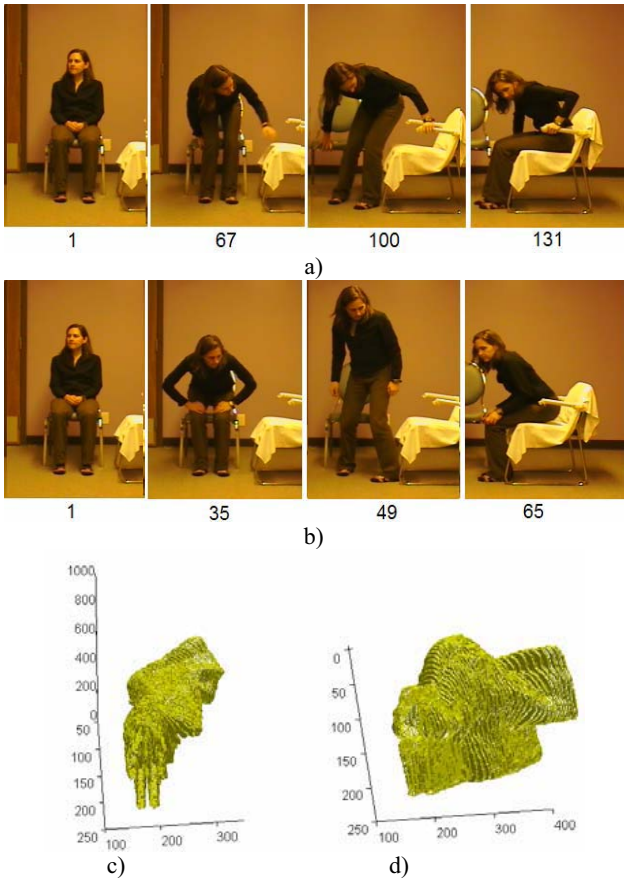

Figure 7. a) key frames in abnormal transfer; b) key frames in normal transfer; c) VMHI for abnormal transfer; d) VMHI for normal transfer.

Table 2. Statistical variance of the unit normals to VMHI surfaces

Activity	Start- end frames	var (n_x)	var (n_y)	var (n_z)
Sit to stand abnormal	1-84	0.3430	0.3373	0.3183
Sit to stand normal	1-56	0.2778	0.1912	0.5249
Stand to sit abnormal	1-90	0.2239	0.4583	0.3311
Stand-to-sit normal	1-60	0.2108	0.3086	0.4765
Reaching abnormal	102-173	0.4446	0.3477	0.2064
Reaching normal	33-72	0.2987	0.2554	0.4443
Stepping abnormal	1-98	0.3898	0.2099	0.3950
Stepping normal	1-30	0.3392	0.1819	0.4732
Picking up abnormal	1-200	0.4147	0.2872	0.2972
Picking up normal	1-50	0.3093	0.1866	0.4945
Transfer abnormal	1-102	0.3743	0.2225	0.3973
Transfer normal	1-65	0.3019	0.1486	0.5287

Table 3. Deviation from the normal pattern of activity

Activity	D_x (%)	D_y (%)	Dominant motion irregularity
Sit-to-stand	23.47	76.41	Horizontal sway
Stand-to-sit	1.31	48.51	Horizontal sway
Reaching	48.84	36.13	Vertical sway
Stepping	14.91	10.99	Temporary stop
Picking up	34.07	53.62	Horizontal sway

C. Speed analysis

Every slice in the VMHI representation as defined by Eq. (1) contains quantitative information about the motion that occurs in the corresponding pair of adjacent frames. Hence, the speed of movement S is computed on a frame by frame basis for the entire video sequence of length N as follows:

$$S(k) = card(VMHI(:,:,k)) \quad k = 1..N \tag{3}$$

where *card* denotes the cardinal of the set represented by the binary VMHI slice. In other words, speed is measured as the number of pixels in motion between each pair of adjacent frames. This simple measure serves well the purpose of this study, which deals with a global assessment of the irregularities in motion. Future work will focus on assessing the relative speed of motion of body parts, which will allow for the characterization of more subtle, local irregularities.

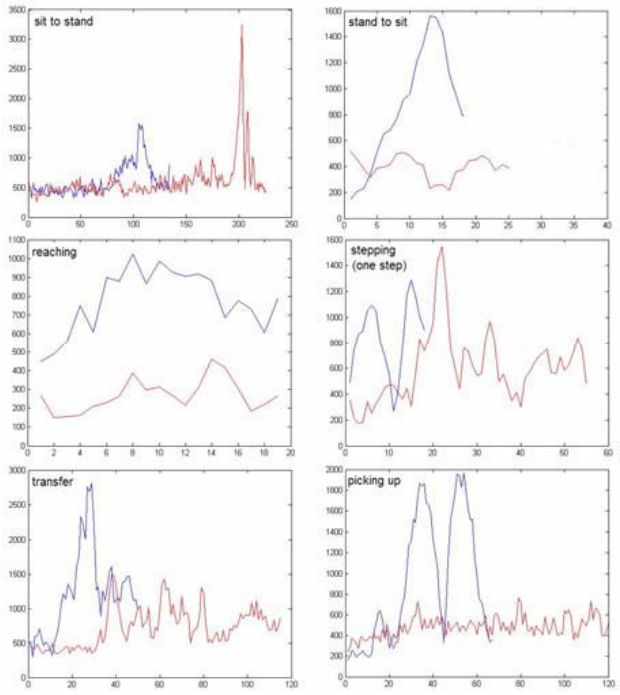

Figure 8. Speed plots for abnormal and normal activities. Red encodes abnormal motion, while blue encodes normal motion. Horizontal axis is graded in frames and shows time. Vertical axis is graded in no of pixels per frame and shows speed as a function of time.

Figure 8 contains the representation of speed as a function of time for normal and abnormal motion over our database of human activities. These plots allow for the following qualitative observations:

- In most activities, the speed of the normal motion is much higher than the speed of abnormal motion. One exception is abnormal sit-to-stand, where a high acceleration can be noticed towards the end of the motion. This acceleration corresponds to the subject ending her motion by "falling" in her chair instead of reclining smoothly. A similar situation is observed in abnormal stepping, where the subject is unable to control the speed of motion.
- The pattern of accelerations and decelerations occurring in normal motion is quite different from the pattern occurring in abnormal motion. Specifically, abnormal motions such as stepping, transfer, reaching, and picking up contain hesitations, therefore they exhibit a lot more accelerations and decelerations than their normal equivalent.
- The only motion where the patterns of normal and abnormal speed are similar up to an additive constant is reaching. In reaching, the main abnormality is the limited range of motion; the subject finishes her motion once her maximal range of motion is reached. Hence, a postural analysis based on pose estimation would help towards a more precise quantification of abnormalities. This is one of our main future work directions.

V. CONCLUSIONS

This paper proposes a new spatiotemporal template based on the 3D extension of the MHI concept introduced by Aaron and Bobick [8]. Our template handles issues such as motion self-occlusion, variable speed and sequence length in the context of human motion analysis. Consequently, the VMHI representation is suitable for the visualization and quantification of several types of motion irregularities. This work focuses on the analysis of sway and speed irregularities. Horizontal and vertical sways are visualized and quantified via an interactive user interface using the deviation vector, which is a measure of spatiotemporal surface smoothness. The experimental results show that this measure is reliable for quantifying the deviation of the abnormal motion from its corresponding normal motion.

A simple global measure of speed allows for visualizing differences in speed patterns in normal and abnormal motion.

Ongoing work focuses on refining our speed measure in order to be able to extract information about the relative speed of motion of various body parts. We are also interested in developing methods for analysis and quantification of other types of motion irregularities, such as postural changes.

Future work will be focused on the quantitative performance evaluation for motions of different degrees of abnormality, in order to be able to correlate the evolution of elderly subjects over a period of time with customized rehabilitation programs, medication etc.

REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Kruger,, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, pp. 90–126, 2006.

[2] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, pp.68–73, August 2000.

[3] E. Chang and Y.F. Wang, "Introduction to the special section on video surveillance," *Multimedia Systems*, vol. 10, pp.116-117, 2004.

[4] www.eyetoy.com

[5] M. Betke, J. Gips, and P. Flemming, "The Camera Mouse: Visual tracking of body features to provide computer access for people with severe disabilities", *IEEE Trans. on neural systems and rehabilitation eng.*, vol. 10, pp. 1-9, 2002.

[6] J. Cutting and L. Kozlowski, "Recognizing friends by their walk: gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353-356, 1977.

[7] K. Berg, S. Wood-Dauphinee, and J.I. Williams, "The Balance Scale: reliability assessment for elderly residents and patients with an acute stroke," *Scandinavian Journal of Rehabilitation Medicine*, vol. 27, pp 27-36, 1995.

[8] J.K. Aggarwal and Q. Cai, "Human motion analysis: a review", *Computer Vision and Image Understanding (CVIU)*, vol. 73, pp. 428-440, 1999.

[9] T. B. Moeslund and E. Granum, "A survey on computer vision-based human motion capture" *Computer Vision and Image Understanding (CVIU)*, vol. 81, pp. 231-268, 2001.

[10] R.C. Nelson and R. Polana, "Qualitative recognition of motion using temporal texture", *CVGIP: Image understanding*, vol. 56, pp. 78-89, 1992.
[11] R. Polana, R.C. Nelson, "Detection and recognition of periodic, non-rigid motion", *International Journal of Computer Vision*, vol. 23, pp. 261-282, 1997.
[12] R. Cutler and L.S. Davis, "Robust real-time periodic motion detection, analysis, and applications", *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 2, pp. 781–96, 2000.
[13] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates", *IEEE Trans. on Patt. Anal. and Machine Intell*., vol. 23, pp. 257–267, 2001.
[14] J.W. Davis, "Hierarchical motion history images for recognizing human motion", Proc. of IEEE Workshop on Detection and Recognition of Events in Video (EVENT'01), July 2001.
[15] M. Valstar, M. Pantic, and I. Patras, "Motion History for Facial Action Detection in Video", Proc. of IEEE Int'l Conf. on Systems, Man and Cybernetics, pp.635-640, Oct. 2004.
[16] D. Weinland, R. Ronfard, and E. Boyer, "Motion history volumes for viewpoint action recognition", IEEE International Workshop on modeling People and Human Interaction (PHI'05), 2005.
[17] A. Yilmaz and M.Shah, "Actions as objects: a novel action representation", Proc. of IEEE Conf. on Computer Vision Pattern Recognition (CVPR), 2005.
[18] M. Blank, L. Gorelick, E. Schechtman, M. Irani, R. Basri,"Actions as space-time shapes," Proc. of IEEE Conf. on Computer Vision (ICCV), 2005.
[19] A. Bobick, S. Intille, J. Davis, F. Baird, L. Campbell, Y. Ivanov, C. Pinhanez, A. Schutte, and A.Wilson, "The KidsRoom: action recognition in an interactive story environment", Presence, 8(4):367–391, 1999.
[20] A. Branzan Albu, T. Beugeling, N. Virji-Babul, and C. Beach," Analysis of Irregularities in Human Actions with Volumetric Motion History Images", in Proc. of IEEE Workshop on Motion and Video Computing, Austin, February 2007.

Alexandra Branzan Albu received the Ph.D. from the Polytechnic Institute of Bucharest in 2000. In 2001, she joined the Computer Vision and Systems Laboratory at Laval University (QC, Canada) as a postdoctoral researcher and became an Assistant Professor at Laval in 2003. In 2005, she joined the ECE department at the University of Victoria (BC, Canada). Her research interests include computer vision-based human motion analysis and medical imaging. Dr. Branzan Albu is a professional engineer affiliated to the Province of British Columbia Association of Professional Engineers. She is also the Technical Program Chair of the IEEE Victoria Section.

Trevor Beugeling graduated from Simon Fraser University (BC, Canada) as a Bachelor in Science (Pure Mathematics) in May 2002. He is currently working towards graduating from the B. Eng. Program in Computer Engineering at the University of Victoria. After his graduation, expected in April 2008, he intends to pursue graduate studies in Computer Vision. He has spent three NSERC-sponsored research terms working in Dr. Branzan Albu's computer vision lab on projects related to human motion analysis and medical imaging. His interests include computer vision, cryptography, and signal coding.