HMM Based Falling Person Detection Using Both Audio and Video*

B. Uğur Töreyin[1], Yiğithan Dedeoğlu[2], A. Enis Çetin[1]

[1] Department of Electrical and Electronics Engineering,
[2] Department of Computer Engineering,
Bilkent University 06800 Bilkent, Ankara, Turkey
{bugur, yigithan, cetin}@bilkent.edu.tr

Abstract. Automatic detection of a falling person in video is an important problem with applications in security and safety areas including supportive home environments and CCTV surveillance systems. Human motion in video is modeled using Hidden Markov Models (HMM) in this paper. In addition, the audio track of the video is also used to distinguish a person simply sitting on a floor from a person stumbling and falling. Most video recording systems have the capability of recording audio as well and the impact sound of a falling person is also available as an additional clue. Audio channel data based decision is also reached using HMMs and fused with results of HMMs modeling the video data to reach a final decision.

1 Introduction

Detection of a falling person in an unsupervised area is a practical problem with applications in safety and security areas including supportive home environments and CCTV surveillance systems. Intelligent homes will have the capability of monitoring activities of their occupants and automatically provide assistance to elderly people and young children using a multitude of sensors including surveillance cameras in the near future [1, 2, 3]. Currently used worn sensors include passive infrared sensors, accelerometers and pressure pads. However, they may produce false alarms and elderly people simply forget wearing them very often. Computer vision based systems propose non-invasive alternatives for fall detection. In this paper, a video based falling person detection method is described. Both audio and video tracks of the video are used to reach a decision.

Video analysis algorithm starts with moving region detection in the current image. Bounding box of the moving region is determined and parameters describing the bounding box are estimated. In this way, a time-series signal describing the motion of a person in video is extracted. The wavelet transform of this signal is computed and used in Hidden Markov Models (HMMs) which were trained according to possible human being motions. It is observed that the wavelet transform domain signal

* This work is supported in part by European Commission 6th Framework Program with grant number FP6-507752 (MUSCLE Network of Excellence Project).

provides better results than the time-domain signal because wavelets capture sudden changes in the signal and ignore stationary parts of the signal.

Audio analysis algorithm also uses the wavelet domain data. HMMs describing the regular motion of a person and a falling person were used to reach a decision and fused with results of HMMs modeling the video data to reach a final decision.

In [4] motion trajectories extracted from an omnidirectional video are used to determine falling persons. When a low cost standard camera is used instead of an omnidirectional camera it is hard to estimate moving object trajectories in a room. Our fall detection method can be also used together with [4] to achieve a very robust system, if an omnidirectional camera is available. Another trajectory based human activity detection work is presented in [5]. Neither [4] nor [5] used audio information to understand video events.

In Section 2, the video analysis algorithm is described and in Section 3, the audio analysis algorithm is presented. In Section 4, experimental results are presented.

2 Analysis of Video Track Data

Our video analysis consists of three steps: i) moving region detection in video, ii) calculation of wavelet coefficients of a parameter related with the aspect ratio of the bounding box of the moving region, and iii) HMM based classification using the wavelet domain data. Each step of our video analysis algorithm is explained in detail next.

i) Moving region detection: The camera monitoring the room is assumed to be stationary. Moving pixels and regions in the video are determined by using a background estimation method developed in [6]. In this method, a background image B_{n+1} at time instant $n+1$ is recursively estimated from the image frame I_n and the background image B_n of the video as follows:

$$B_{n+1}(k,l) = \begin{cases} aB_n(k,l) + (1-a)\, I_n(k,l), & \text{if } I_n(k,l) \text{ stationary} \\ B_n(k,l), & \text{if } I_n(k,l) \text{ moving} \end{cases} \quad (1)$$

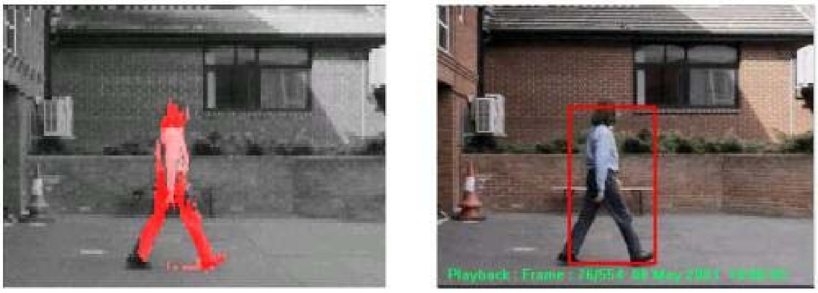
where $I_n(k, l)$ represents a pixel in the n^{th} video frame I_n, and a is a parameter between 0 and 1. Moving pixels are determined by subtracting the current image from the background image and adaptive thresholding (cf. Fig. 1a). For each pixel an adaptive threshold is estimated recursively in [6]. Pixels exceeding thresholds form moving regions and they are determined by connected component analysis.

We do not need very accurate boundaries of moving regions. Hence the above computationally efficient algorithm is sufficient for our purpose of estimating the aspect ratios of moving regions in video. Other methods including the ones described in [7] and [8] can also be used for moving pixel estimation but they are computationally more expensive than [6].

ii) Feature extraction from moving regions and the wavelet transform: After a post-processing stage comprising of connecting the pixels, moving regions are encapsulated with their minimum bounding rectangles (cf. Fig.1b). Next, the aspect ratio, ρ, for each moving object is calculated. The aspect ratio of the i^{th} moving object is defined as:

$$\rho_i(n) = \frac{H_i(n)}{W_i(n)} \qquad (2)$$

where $H_i(n)$ and $W_i(n)$ are the height and the width of the minimum bounding box of the i^{th} object at image frame n, respectively, We then calculate the corresponding wavelet coefficients for ρ. Wavelet coefficients, w_i's, are obtained by high-pass filtering followed by decimation as shown in Fig. 2.



(a) **(b)**

Fig. 1. **(a)** Moving pixels, and **(b)** their minimum bounding boxes are determined

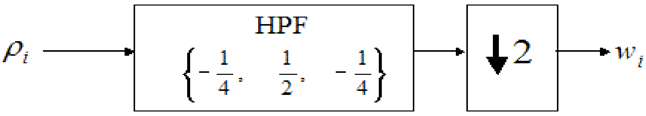


Fig. 2. Wavelet coefficients, w_i corresponding to aspect ratio ρ_i are evaluated with an integer arithmetic high-pass filter (HPF) corresponding to Lagrange wavelets [9] followed by decimation

The wavelet transform of the one-dimensional aspect ratio signal is used as a feature signal in HMM based classification in this paper. It is experimentally observed that the aspect ratio based feature signal exhibits different behaviour for the cases of walking and falling persons. A quasi-periodic behaviour is obviously apparent for a walking person in both $\rho(n)$ and its corresponding wavelet signal as shown in Fig. 3. On the other hand, the periodic behaviour abruptly ends and $\rho(n)$ decays to zero for a falling person or a person sitting down. This decrease and the later stationary characteristic for fall is also apparent in the corresponding subband signal (cf. Fig. 4).

Using wavelet coefficients, w, instead of aspect ratios, ρ, to characterize moving regions has two major advantages. The primary advantage is that, wavelet signals can easily reveal the aperiodic characteristic which is intrinsic in the falling case. After the fall, the aspect ratio does not change or changes slowly. Since, wavelet signals are high-pass filtered signals, slow variations in the original signal lead to zero-mean wavelet signals. Hence it is easier to set thresholds in the wavelet domain which are robust to variations of posture sizes and aspect ratios for different people. This constitutes the second major advantage. We set two threshols, *T1* and *T2* for defining Markov states in the wavelet domain as shown in Fig. 3. The lower threshold *T1* basically determines the wavelet signal being close to zero. After the fall, ideally the wavelet signal should be zero but due to noise and slow movements of the fallen person the wavelet coefficients wiggle around zero. The use of wavelet domain information also makes the method robust to variations in object sizes. This is achieved by the use of the second threshold *T2* to detect high amplitude variations in the wavelet signal, which correspond to edges or high-frequency changes in the original signal. When the wavelet coefficients exceed the higher threshold *T2* in a frequent manner this means that the object is changing its shape or exhibiting periodic behaviour due to walking or running.

iii) HMM based classification: Two three-state Markov models are used to classify the motion of a person in this paper. Non-negative thresholds *T1* < *T2* introduced in wavelet domain, define the three states of the Hidden Markov Models for walking and falling, as shown in Fig. 5a and b, respectively.

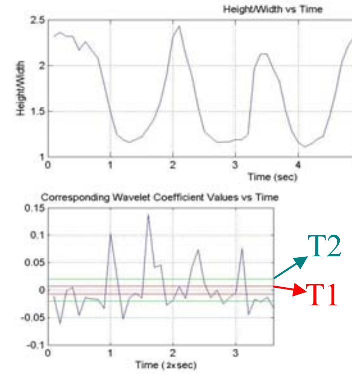


Fig. 3. Quasi-periodic behaviour in ρ vs. time (top), and the corresponding wavelet coefficients w vs. time for a walking person (sampling period is half of the original rate in the wavelet plot). Thresholds *T1* and *T2* introduced in the wavelet domain are robust to variations in posture sizes and aspect ratios of different people

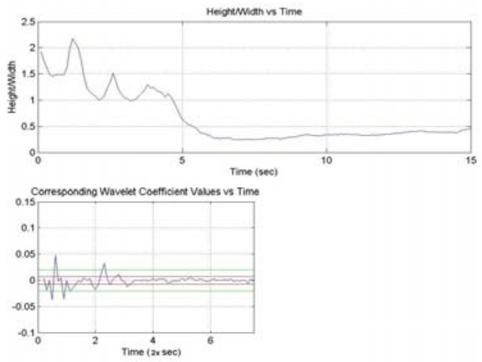
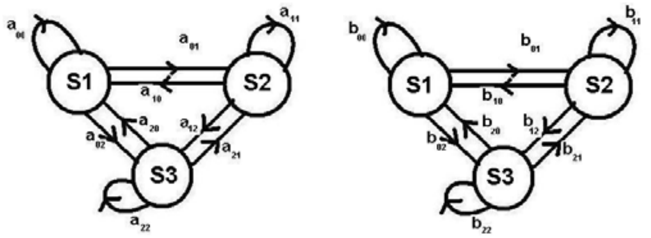


Fig. 4. Aspect ratio ρ vs. time (top), and the corresponding wavelet coefficients w vs. time for a falling person (sampling period is half of the original rate in the wavelet plot)

At time *n,* if $|w_i(n)| < T1$, the state is in S1; if $T1 < |w_i(n)| < T2,$ the state is S2; else if $|w_i(n)| > T2$, the state S3 is attained. During the training phase of the HMMs transition probabilities a_{uv} and b_{uv}, $u,v = 1, 2, 3,$ for walking and falling models are estimated off-line, from a set of training videos. In our experiments, 20 consecutive image frames are used for training HMMs.



(a) **(b)**

Fig. 5. Three state Markov models for **(a)** walking, and **(b)** falling

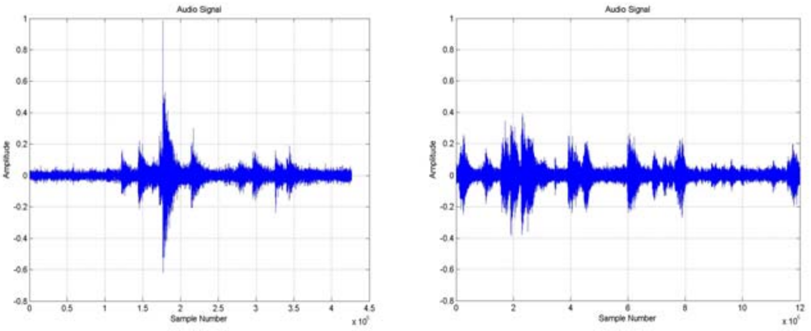
For the walking person, since the motion is quasi-periodic, we expect similar transition probabilities between the states. Therefore the values of *a's* are close to each other. However, when the person falls down, the wavelet signal starts to take values around zero. Hence we expect a higher probability value for b_{00} than any other *b* value in the falling model, which corresponds to higher probability of being in *S1*. The state *S2* provides hysteresis and it prevents sudden transitions from *S1* to *S3* or vice versa.

During the recognition phase the state history of length 20 image frames are determined for the moving object detected in the viewing range of the camera. This state sequence is fed to the walking and falling models. The model yielding higher probability is determined as the result of the analysis for video track data. However, this is not enough to reach a final decision of a fall. Similar *w* vs. time characteristics are observed for both falling and ordinary sitting down cases. A person may simply sit down and stay stationary for a while. To differentiate between the two cases, we incorporate the analysis of the audio track data to the decision process.

3 Analysis of Audio Track Data

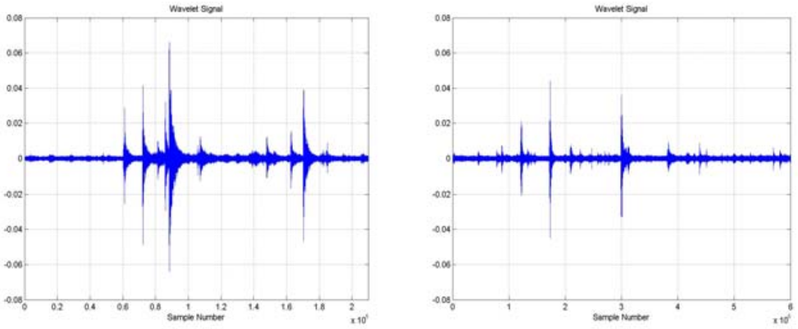
In this paper, audio signals are used to discriminate between falling and sitting down cases. A typical stumble and fall produces high amplitude sounds as shown in Fig. 6a, whereas the ordinary actions of bending or sitting down has no distinguishable sound from the background (cf. Fig. 6b). The wavelet coefficients of a fall sound are also different from bending or sitting down as shown in Fig. 7. Similar to the motivation in the analysis of video track data for using wavelet coefficients, we base our audio analysis on wavelet domain signals. Our previous experience in speech recognition indicates that wavelet domain feature extraction produces more robust results than Fourier domain feature extraction [10]. Our audio analysis algorithm also consists of three steps: i) computation of the wavelet signal, ii) feature extraction of the wavelet signal, and iii) HMM based classification using wavelet domain features.

i) Wavelet signal: We use the same high-pass filter followed by a decimation block shown in Fig. 2, to obtain a wavelet signal corresponding to the audio signal accompanying the video track data. The wavelet signals corresponding to the audio track data in Fig. 6a and b, are shown in Fig. 7a and b, respectively.



(a) **(b)**

Fig. 6. Audio signals corresponding to **(a)** a fall, which takes place at around sample number $1.8\text{x}10^5$, and **(b)** talking (0 - $4.8\text{x}10^5$), bending ($4.8\text{x}10^5$ – $5.8\text{x}10^5$), talking ($5.8\text{x}10^5$ – $8.9\text{x}10^5$), walking ($8.9\text{x}10^5$ – $10.1\text{x}10^5$), bending ($10.1\text{x}10^5$ - $11\text{x}10^5$), and talking ($11\text{x}10^5$ - 12 $\text{x}10^5$) cases. The sound signals are sampled with 44,100 Hz



(a) **(b)**

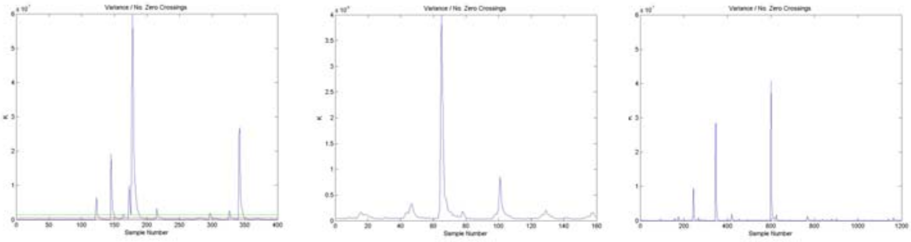
Fig. 7. The wavelet signals corresponding to the audio signals in **(a)** falling (0.9 $\text{x}10^5$), and **(b)** talking (0 – $2.4\text{x}10^5$), bending ($2.4\text{x}10^5$ – $2.9\text{x}10^5$), talking ($2.9\text{x}10^5$ – $4.5\text{x}10^5$), walking ($4.5\text{x}10^5$ – $5\text{x}10^5$), bending ($5\text{x}10^5$ – $5.5\text{x}10^5$), and talking ($5.5\text{x}10^5$ - 6 $\text{x}10^5$).

ii) Analysis of wavelet signals: The wavelet signals corresponding to audio track data are further analyzed to extract features in fixed length short-time windows. We take 500-sample-windows in our implementation. Our sampling frequency is 44.1 KHz. We determine the variance, σ_i^2, and the number of zero crossings, Z_i, in each window i.

We observe that, walking is a quasi-periodic sound in terms of σ_i^2 and Z_i. However, when a person stumbles and falls, Z_i decreases whereas σ_i^2 increases. So we define a feature parameter κ in each window as follows:

$$\kappa = \frac{\sigma^2}{Z_i} \tag{3}$$

where the index i indicates the window number. The parameter κ takes non-negative values.

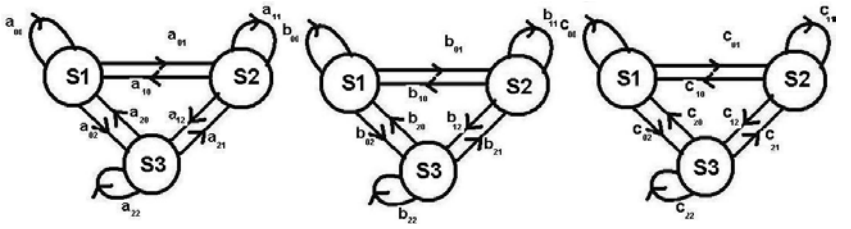


(a) **(b)** **(c)**

Fig. 8. The ratio of variance over number of zero crossings, κ, variations for **(a)** falling (180), **(b)** walking, and **(c)** talking (0 – 480), bending (480 – 590), talking (590 – 900), walking (900 – 1000), bending (1000 – 1100), and talking (1100 - 1200). Note that, κ values for (b) the walking case, are an order of magnitude less than (a) falling and (c) talking cases. Thresholds *T1' < T2'*, are defined in κ domain

Talking has a varying σ_i^2-Z_i characteristic depending on the utterance. When vowels are uttered, σ_i^2 increases while Z_i decreases, which results in larger κ values compared to consonant utterances. Variation of κ values versus sample numbers for different cases, are shown in Fig. 8.

iii) HMM based classification: In this case, three three-state Markov models are used to classify the walking, talking and falling sounds. The non-negative thresholds *T1' < T2'* introduced in κ domain, define three states of the Hidden Markov Models for walking, talking and falling, as shown in Fig. 9a, b, and c, respectively.



(a) **(b)** **(c)**

Fig. 9. Three-state Markov models for **(a)** walking, **(b)** talking, and **(c)** falling sound classification

For the i^{th} window of the wavelet signal, if $|\kappa_i| < T1'$, state S1; if $T1' < |\kappa_i| < T2'$, state S2; else if $|\kappa_i| > T2'$, state S3 is attained. During the training phase, transition probabilities, a_{uv}, b_{uv}, and c_{uv}, $u,v = 1, 2, 3,$ for walking, talking, and falling models, respectively, are estimated off-line. These probabilities are estimated from 20 consecutive κ values corresponding to 20 consecutive 500-sample-long wavelet windows for training HMMs.

During the classification phase a state history signal consisting of 20 κ values are estimated from the sound track of the video. This state sequence is fed to the walking, talking, and falling models in running windows. The model yielding highest probability is determined as the result of the analysis for audio track data. We then combine this result with the result of the video track analysis step using the logical "and" operation. Therefore, a "falling person detected" alarm is issued only when both video and audio track data yield the highest probability in their "fall" models.

4 Experimental Results

The proposed algorithm works in real-time on an AMD AthlonXP 2000+ 1.66GHz processor. As described above HMMs are trained from falling, walking, and walking and talking video clips. A total of 64 video clips having 15,823 image frames are used. In all of the clips, only one moving object exists in the scene. Contents of the test video clips are summarized in Table 1.

Table 1. Video content distribution in the test set

Video Content	Include Audio	No. of Clips
Walking + Talking	Yes	16
Sitting down + Talking	Yes	5
Sitting down	Yes	4
Walking + Falling	Yes	25
Walking + Falling	No	14

As can be seen from Table 1, 14 of the clips having falls do not have audio track data, hence we only make use of the video track data analysis part of our method to determine whether falling takes place. Image frames from the above video clips are shown in Figure 10.

The classification results for the above test data with only video analysis and both audio and video analysis are presented in Table 2. There is no way to distinguish a person intentionally sitting down on the floor from a falling person, if only video track data is used. When both modalities are utilized, they can be distinguished and we do not get any false positive alarms for the videos having a person sitting down as shown in Table 2.

Table 2. Detection results for the test set

Video Content	Include Audio	No. of Clips	No. of Clips in which Falling is Detected	
			Video	Audio+Video
Walking + Talking	Yes	16	0	0
Sitting down + Talking	Yes	5	5	0
Sitting down	Yes	4	4	0
Walking + Falling	Yes	25	25	25
Walking + Falling	No	14	14	14

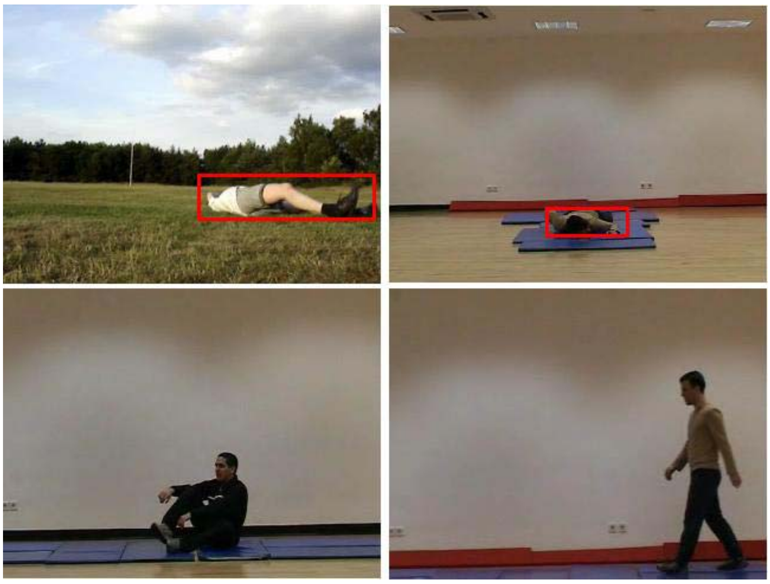


Fig. 10. Image frames from falling, sitting, and walking and talking clips

5 Conclusion

A method for automatic detection of a falling person in video is developed. Main contribution of this work is the use of both audio and video tracks to decide a fall in video. The audio information is essential to distinguish a falling person from a person simply sitting down or sitting on a floor. Three-state HMMs are used to classify events. Feature parameters of HMMs are extracted from temporal wavelet signals describing the bounding box of moving objects. Since wavelet signals are zero-mean signals, it is easier to define states in HMMs and this leads to a robust method against variations in object sizes.

The method is computationally efficient and it can be implemented in real-time in a PC type computer.

Similar HMM structures can be also used for automatic detection of accidents and stopped vehicles in highways which are all examples of instantaneous events occurring in video.

References

1. Barnes, N.M., Edwards, N.H., Rose, D.A.D., Garner, P.: Lifestyle Monitoring: Technology for Supported Independence. IEE Comp. and Control Eng. J. (1998) 169-174
2. Bonner, S.: Assisted Interactive Dwelling House: Edinvar Housing Assoc. Smart Tech. Demonstrator and Evaluation Site In: Improving the Quality of Life for the European Citizen (TIDE), (1997) 396–400
3. McKenna, S.J., Marquis-Faulkes, F., Gregor, P., Newell, A.F.:Scenario-based Drama as a Tool for Investigating User Requirements with Application to Home Monitoring for Elderly People. In Proc. of HCI, (2003)
4. Nait-Charif, H., McKenna, S.: Activity Summarisation and Fall Detection in a Supportive Home Environment. In Proc. of ICPR'04, (2004) 323-326
5. Cuntoor, N.P., Yegnanarayana, B., Chellappa, R.: Interpretation of State Sequences in HMM for Activity Representation. In Proc. of IEEE ICASSP'05, (2005) 709-712
6. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L.: A System for Video Surveillance and Monitoring: VSAM Final Report.Tech. Report CMU-RI-TR-00- 12, Carnegie Mellon University (1998)
7. Bagci, M., Yardimci, Y., Cetin, A.E.:Moving Object Detection Using Adaptive Subband Decomposition and Fractional Lower Order Statistics in Video Sequences. Elsevier, Signal Processing. (2002) 1941—1947
8. Stauffer, C., Grimson, W.E.L.: Adaptive Background Mixture Models for Real-Time Tracking. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (1999) 246-252
9. Kim, C.W., Ansari, R., Cetin, A.E.: A class of linear-phase regular biorthogonal wavelets. In Proc. of IEEE ICASSP'92 (1992) 673-676
10. Jabloun, F., Cetin, A.E., Erzin, E.: Teager Energy Based Feature Parameters for Speech Recognition in Car Noise. IEEE Signal Processing Letters (1999) 259-261