Moving Humans Detection Based on Multi-modal Sensor Fusion

Bir Bhanu and Xiaotao Zou

Center for Research in Intelligent Systems
University of California, Riverside, California, 92521, USA
{bhanu, xzou}@vislab.ucr.edu

ABSTRACT

Moving object detection plays an important role in automated surveillance systems. However, it is challenging to detect moving objects robustly in a cluttered environment. In this paper, we propose an approach for detecting humans using multi-modal measurements. The approach is based on using Time-Delay Neural Network (TDNN) to fuse the audio and video data at the feature level for detecting the walker with multiple persons in the scene. The main contribution of this paper is the introduction of Time-Delay Neural Network in learning the relation between visual motion and step sounds of the walking person. Experimental results are presented.

1. INTRODUCTION

Automated surveillance addresses real-time observation of people and vehicles within a busy environment, leading to a description of their actions and interactions. The technical issues include moving object detection and tracking, object classification, human motion analysis, activity understanding, etc. Most commonly used sensors for surveillance are imaging sensors, e.g. video cameras and thermal imaging systems. There are a number of video surveillance systems [1] [2] [3], which consist of a single camera or hundreds of cameras. To achieve the 24-7 continuous monitoring, some IR cameras are used along with the optical cameras under low illuminations [4].

Besides these video or IR surveillance systems, there also exist "detection and tracking" systems based on non-imaging measurements. A project named "Smart Floor" [5] aims to identify and track a user around the space with force measuring load cells installed under the floor. However, along with the relatively high performance comes the high cost and careful design of the instrumented space. In most common scenarios a simple sensor network composed of off-the-shelf video cameras and microphones may be appropriate. Since the audio-video sequences cover overlapping areas, a sensor fusion mechanism should be used to obtain a more accurate and more efficient solution.

The surveillance task can be described as the sequential processes where multiple moving objects are to be extracted (detection), followed (tracking), distinguished (recognition), and their interactions understood (activity recognition) [4]. Each part of this process presents challenging problems and may depend on the outcome of the previous step. At the core of this process is the detection module. This paper addresses the issue of detecting walking persons on the basis of audio-video measurements.

2. RELATED WORK AND MOTIVATION

2.1. Related work

Several studies on fusing audio-video data for detection have been reported in the literature. Fisher et al. [6] presented an information theoretic approach for the fusion of multiple modalities, which can detect where a speaker is within a scene and whether he/she is producing specific words. Garg et al. [7] developed a supervised learning framework based on Bayesian networks and applied it to the problem of audio/visual speaker detection for a smart kiosk. Stork et al. [8] proposed a modified Time-Delay Neural Network (TDNN) to perform the visual lip-reading to improve the accuracy of acoustic speech recognition. Cutler and Davis [9] also used the TDNN to learn the audio/visual correlation for searching the speaking person in the scene. However, in all the works mentioned above, it is assumed that the objects within the scene (e.g., the speaking faces) do not move dramatically. Consequently, they cannot address the dynamic changes of the scene.

Ikeda et al. [10] presented an approach to fuse multi-modal sensory data for moving object detection (e.g. clapping hands, walkers) based on mutual information maximization. The changes produced by the movement of objects are addressed by tracking centroids of detected regions. To simplify the computation of mutual information, the authors assumed that the input visual/audio signals are jointly Gaussian. However, this statistical assumption weakens the application of this approach in the real world.

This paper explores the relation between visual motions and step sounds, and discusses the application of TDNN in multi-modal sensory fusion for walking human

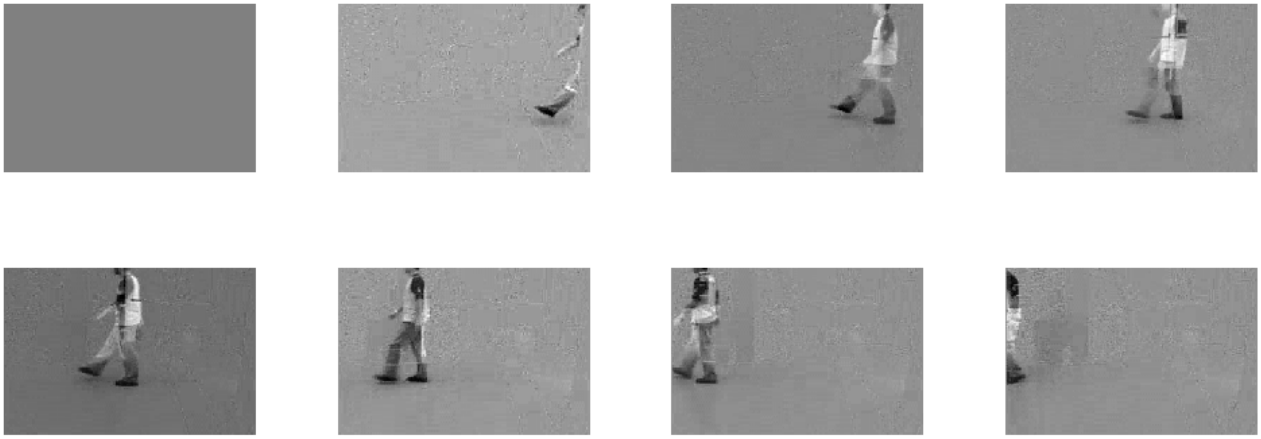

Figure 1. Walking human sequence (extracted from the background)

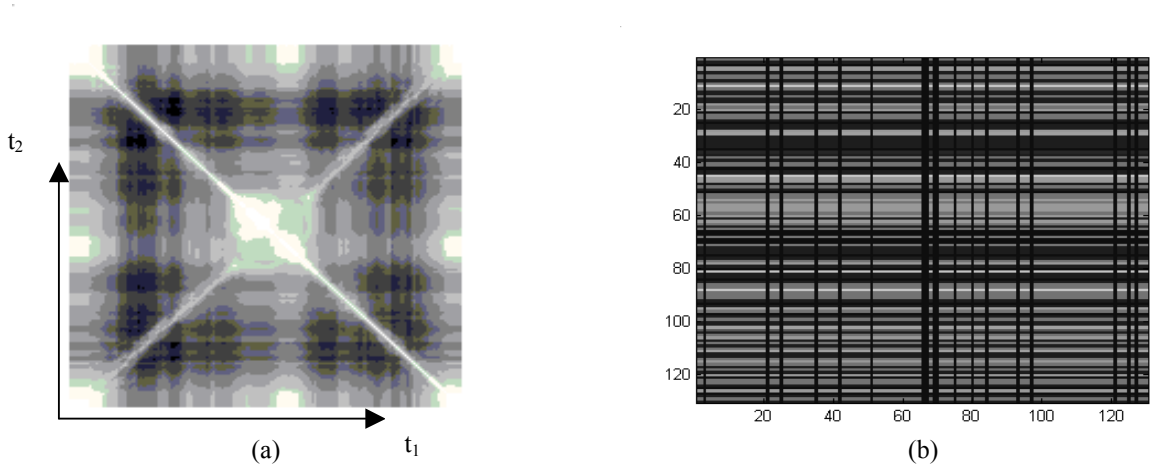

Figure 2. (a) Recurrence matrix of video signals; (b) Similarity of stepping sounds

detection. We present a method for automatically detecting a walking person (both spatially and temporally) by fusing video and audio data. The audio-visual correlation is first learned by a time-delay neural network, which then performs a spatio-temporal search for the walking person.

2.2. Motivation

The visual motion of a walking person (a.k.a. gait, shown in Figure 1) is periodic and highly correlated with the corresponding audio data (step sounds). A similar fact (correlation between the motion of mouths during speaking and the speech sounds) has been exploited for lip-reading [8] and speaker detection [9]. Figure 2 shows a recurrence matrix of the extracted human motion and the similarity of the corresponding audio data. A recurrence matrix is a qualitative tool used to perform time series analysis of non-linear dynamic systems. In this case, the recurrence matrix RM is defined by:

$$RM(t_1, t_2) = R(I_{t_1}, I_{t_2}) \qquad (1)$$

Where R is the correlation function, and I_{t1}, I_{t2} denote frame images at times t_1 and t_2 respectively. Similarly, we use the Euclidean distance between the magnitudes to define the similarity of the corresponding audio signals at times t_1 and t_2.

In Figure 2, we can see that the change in the audio data is highly correlated with visual change in the gait. It prompts us to use some mechanism to detect the walking persons in the scene by fusing these two different modalities of signals, a.k.a. visual and audio.

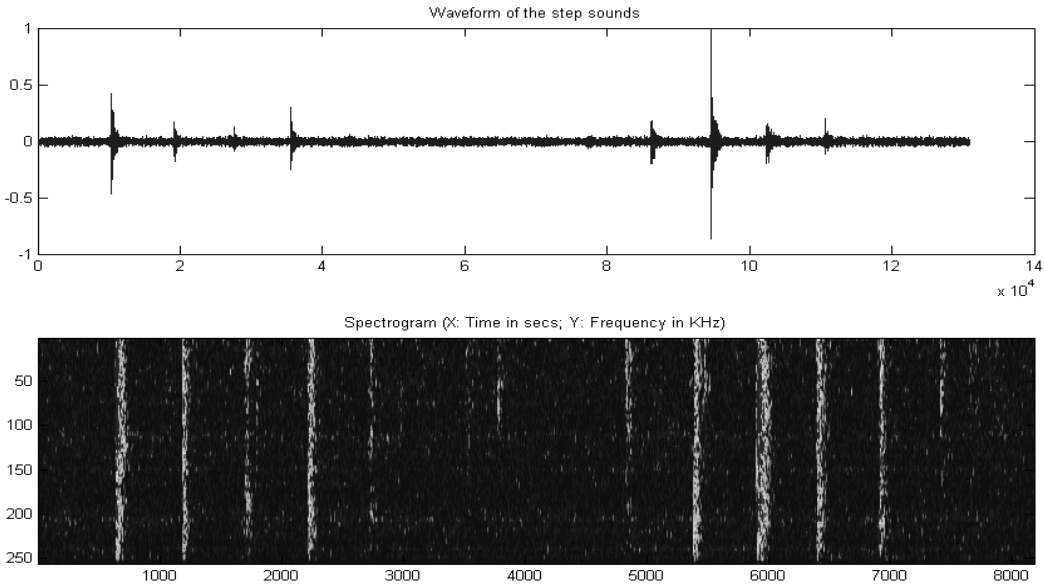

Figure 3. Waveform (upper row) and spectrogram (lower row) of the step sounds

3. TECHNICAL APPROACH

3.1. Preliminary detection

Given the video sequences with a walking person in the scene, we need to first detect and extract all moving objects. Detecting motion is defined as "the ability to cope with moving and changing objects, changing illumination, and changing viewpoints". Based on this definition several methods have been reported. The methods can be roughly viewed as feature-based and featureless methods. In a feature-based method, a set of features, f_i, is detected in each image formed at time t. Features are detected and tracked over a period of time (i.e., over a number of frames). The Featureless methods rely on the pixel values and make no assumptions on the structure in the scene. Two major methods widely used are the Background Subtraction (sometimes called frame differencing) and the Optical Flow. Considering the assumption of relatively fixed environment, we employed the background subtraction algorithm for our preliminary detection.

Background subtraction (or temporal differencing) is conceptually one of the simplest approaches to detecting changes and detecting moving objects. Frame differencing in its simplest definition is when we directly compare the corresponding pixels of the two frames to determine whether they are the same. In its simplest form, the binary difference picture $DP_{jk}(x,y)$ between frames F(x,y,j) and F(x,y,k) is obtained by :

$$DP_{jk}(x,y)=\begin{cases}1, if\ |F(x,y,j)-F(x,y,k)|>\tau\\ 0, otherwise\end{cases} \quad (2)$$

Slowly moving objects and slowly varying intensity changes may not be detected for a given threshold value τ. Using this technique, the difference image will have many noisy pixels. First, a size filter may reduce the noise. This, however, may also filter out some desired signals such as those from slow moving objects. A more robust algorithm may use a local mask, and compare the intensity distributions around the pixel. One may compare the frames using the likelihood ratio:

$$\lambda=\frac{\left[\frac{\sigma_1+\sigma_2}{2}+\left(\frac{\mu_1-\mu_2}{2}\right)^2\right]^2}{\sigma_1*\sigma_2} \quad (3)$$

Where μ and σ denote the mean gray value and variances for the example areas from the frames.

The Difference equation will now be:

$$DP_{jk}(x,y)=\begin{cases}1, if\ \lambda>\tau;\\ 0, otherwise.\end{cases} \quad (4)$$

The background subtraction result is shown in Figure 8.b.

3.2. Feature extraction

We utilize a TDNN to learn the relation between the audio and visual signals. For visual features, we choose a simple measure of change between two images I_{t1} and I_{t2} (i.e. normalized cross-correlation):

$$R_{t,t+i} = \frac{\sum_{(x,y)\in W} (I_t(x,y) - \overline{I_t})(I_{t+i}(x,y) - \overline{I_{t+i}})}{\{\sum_{(x,y)\in W_1} |I_t(x,y) - \overline{I_t}|^2 \sum_{(x,y)\in W_2} |I_{t+i}(x,y) - \overline{I_{t+i}}|^2\}^{1/2}} \quad (5)$$

In the moving object detection application, it will fail if the whole image is used to compute the cross-correlation. For solving this problem, our approach is to track the center of the person. As mentioned above, the objects in the scene (walking and non-walking) have been detected by background subtraction. Prior to fusing video and audio the trajectory of each detected object is computed. By computing correlations between each subtracted objects along the acquired trajectory, we obtain a stable relation over time. In Figure 2.a, our desired visual feature (frame correlations $R_{t,t+1}$) lies on the line immediately below the diagonal.

The sound spectrogram (also called sonogram) coefficients of the step sounds are used as the audio features. The sound spectrogram, like a musical score, is a visual representation of sound. It is acquired by computing a short-time Fourier transform (STFT) of the one-dimensional audio signal. As shown in Figure 3, the horizontal dimension corresponds to time, and the vertical dimension denotes frequency. Frequency is measured in Kilo-Hertz (KHz). The relative intensity of the sound at a particular time and frequency is indicated by the brightness of the spectrogram at that point.

3.3 Time-Delay Neural Network

Figure 4 shows a typical TDNN architecture; while the architecture consists of input, hidden and output layers, much as the classical neural nets, there is a crucial difference. Each hidden unit accepts input from a restricted spatial range of positions in the input layer. Hidden units at "delayed" locations (i.e., shifted to the right) accept inputs from the input layer that are similarly shifted. Training proceeds as in standard back propagation, but with the added constraint that corresponding weights are forced to have the same value - an example of weight sharing. Thus, the weights learned do not depend upon the position of the pattern (as long as the full pattern lies in the domain of the input layer).

In our approach, the TDNN has an input layer consisting of 4 audio features [C_t], which include

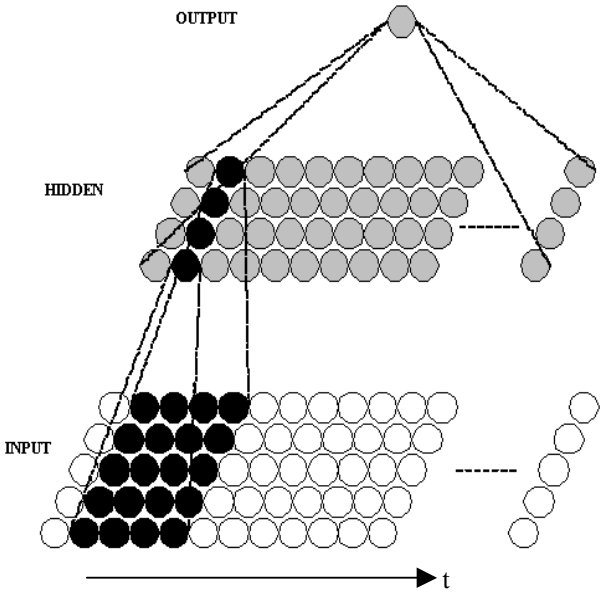

Figure 4. A typical time-delay neural network (TDNN)

spectrogram coefficients at 10, 1K, 10K and 100K Hz, and 1 visual feature [$S_{t,\ t+1}$], i.e. the normalized cross-correlation at each time *t*. In our experiments, we chose the time span such that approximately 1s (time for a common walking cycle) of context is provided. There is one hidden layer with 4 elements at each time *t*, and only a single output O_t, which indicates the possibility of the focused object to be walking at the time *t*.

The TDNN is trained using supervised learning and back propagation. Specifically, for each object P_t, the output O_t is set to 1 where it is exactly walking and 0 otherwise. The training data consists of both positive (O_t=1) and negative (O_t=0) situations. The feedforward operation of the network (during detection) is the same as in standard three-layer networks, but because of the weight sharing, the final output does not depend upon the position of the input. The network gets its name from the fact that it was developed for, and finds the greatest use in speech and other temporal phenomena, where the shift corresponds to delays in time [11]. Once the TDNN has been trained, it is evaluated on an audio-video sequence to detect correlated motion and audio that is indicative of a person walking.

3.4 Dataflow diagram

Finally, the approach described here can be summarized by the following dataflow diagram.

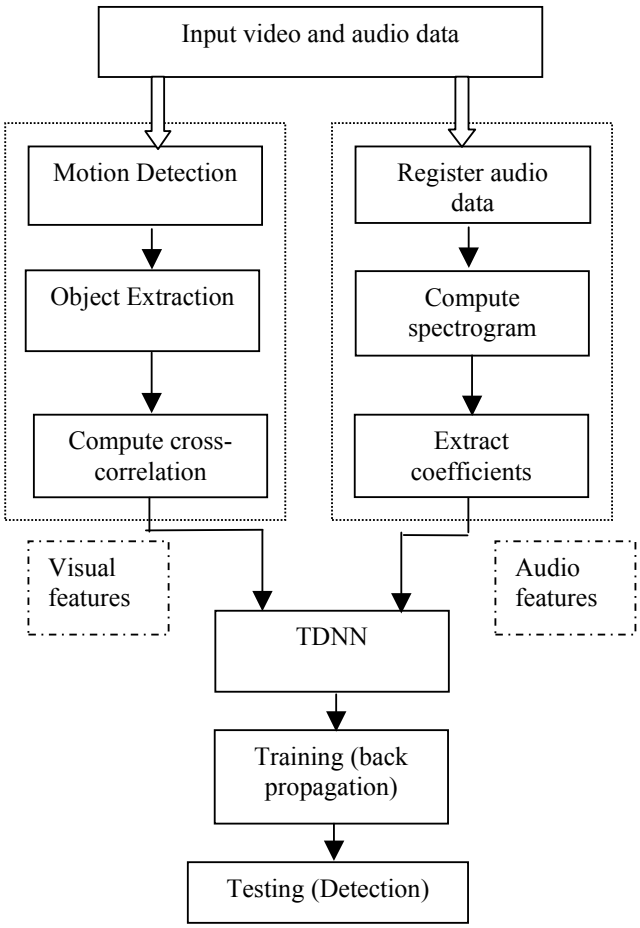

Figure 5. Data Flow of object detection using bi-modal sensor fusion

4. EXPERIMETAL RESULTS

4.1. Assumptions

A scenario is employed in our experiment: there are multiple persons in the indoor environment monitored by a video camera and an off-the-shelf microphone. Specifically, there is only one human walking in the room while the other one is making sort of silent motions to "fool" the detection system. We recorded the video and audio data using a SONY DCR-VX1000 Digital Handycam. It has three 1/3" CCDs with 410,000 pixels for each. The sound was recorded with the built-in microphone and the audio was sampled at 32KHz 12-bit resolution. With the video camera mounted on the fixed tripod, we recorded two sequences (video and audio). The first one is for training, which is shown in Figure 6. The second one is divided into six video clips, all for testing the performance (shown in Figure 7).

4.2. Walker detection results

By the video data alone, we can detect the walking person in the scene. Unfortunately, the waving hands would also be picked out by the motion detection algorithm (shown in Figure 8.c). To address this issue, we fuse the visual motion and step sounds to detect the real walker.

Firstly, we executed the preliminary detection by using the modified Background Subtraction. To initialize the background, we chose the initial frame without any human. After the pre-detection, all humans appearing in the scene are extracted. In our training data ("Test1"), there is only one object in the scene, a.k.a. the walker. In the testing sequences, there are one walker and another false "object", which is fooling the system by waving his hands or turning his body and head (shown in Figure 8.b).

All extracted objects are sent into the Feature Extraction module to compute the visual cross-correlations and audio coefficients. In our experiment, we chose different numbers of features (in Table 1) to test the performance of TDNN. The TDNN has been implemented and evaluated in MATLAB 6.1 on a Pentium 4 1.7 GHz notebook with 256 MB memory.

The performance and time complexity of these settings are shown in Figure 9. It clearly shows the tradeoff between the detection rate and the time complexity. With totally 8 features (4 visuals and 4 audios), all walking objects can be correctly detected. However, the training time also increases exponentially along with the number of features. To achieve the best tradeoff, we chose the third one (#=5) as our default setting.

Figure 6. The training video sequence

(Test sequence #1)

(Test sequence #2)

(Test sequence #3)

(Test sequence #4)

(Test sequence #5)

(Test sequence #6)

Figure 7. The test sequences

(a) 110th frame of 'Test2' (b) BS Detection result (c) Motion Detection result (d) our result

Figure 8. Comparison of Background Subtraction (BS), Motion Detection and Fusion-based Detection

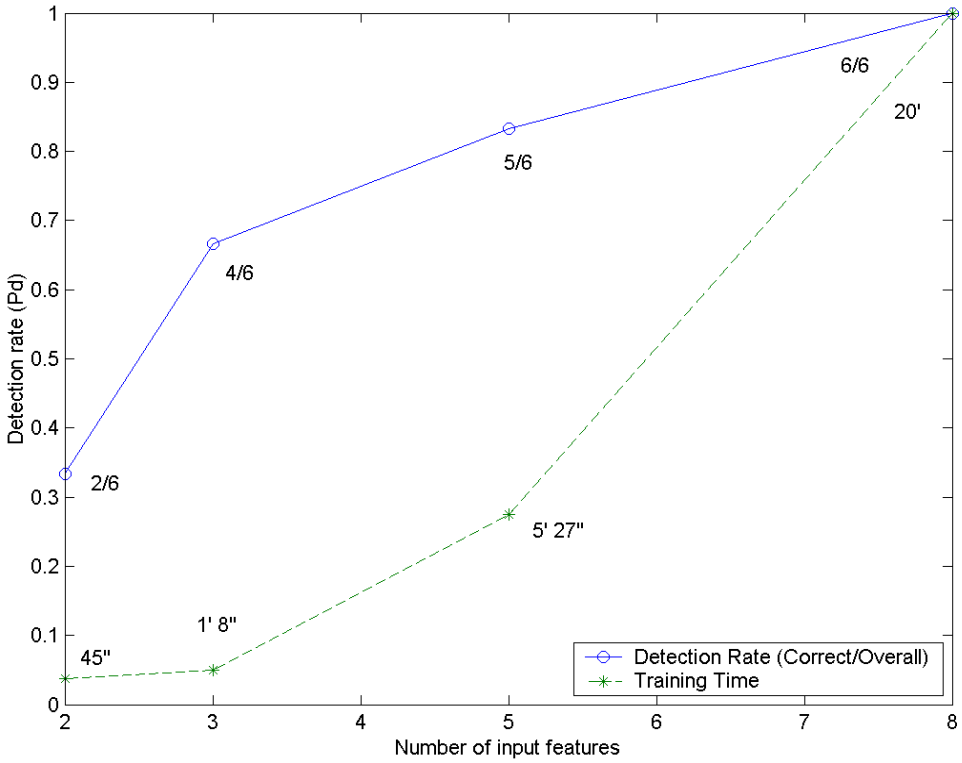

Figure 9. Performance of TDNN with different numbers of input features

After obtaining both visual and audio features from the training sequence "Test1", we fed them into the time-delay neural network described earlier to extract the weights $\{W_1\ W_2 \ldots W_{84}\}$.

Number of features	2*	3**	5	8***
Feature Combination	1 visual + 1 audio	1 visual + 2 audio	1 visual + 4 audio	4 visual + 4 audio

Table 1. Different setting of input features

*: The audio feature is the magnitude of step sounds;
**: The audio features are coefficients at 10 and 10K Hz;
***: The 4 visual features are the correlations $R_{t, t+1}$, $R_{t, t+2}$, $R_{t, t+3}$, $R_{t, t+4}$

Then, we applied the trained network on the test sequence "Test2" (totally 6 clips for testing). After fusing the video and audio data using the TDNN, the true walker is detected from the scene accurately (5 out of 6 totally) with a much lower false alarm rate. The misclassified clip is "Test sequence #6", in which the non-walker shows sort of walking-alike hand movements. The confusion matrix is given below (Table 2).

Ground Truth (GT)	Walkers detected	Non-walkers detected
Walkers (GT)	5	1
Non-walkers (GT)	0	6

Table 2. Confusion matrix

5. CONCLUSIONS

In this paper, we proposed the use of multi-modal sensor fusion for walking human detection. Specifically, we presented an approach for detecting walking humans based on video sequences and step sounds. The proposed method fed the visual cross-correlations and sound spectrogram coefficients to a Time-Delay Neural Network to train and detect the walking humans. It adapts to the dynamic changes of the interested objects and provides the registration between data acquired by sensors of different modalities. Experimental results of walker localization are carried out to confirm the effectiveness of the proposed method.

6. REFERNECES

[1] R. T. Collins, A. J. Lipton, H. Fujiyoshi, T. Kanade, "Algorithms for cooperative multisensor surveillance", *Proc. IEEE*, Vol. 89, No. 10, pp. 1456-1477, Oct. 2001;

[2] J. C. Cheng, J. M. F. Monra, "Model-based recognition of human walking in dynamic scenes", *Proc. IEEE 1st Workshop on Multimedia Signal Processing*, pp. 263-273, 1997;

[3] J. M. Ferryman, S. J. Maybank, A. Worrall, "Visual surveillance for moving vehicles", *Intl. J. Compu. Vis*., Vol. 37, No. 2, pp. 187-197, Oct. 1998;

[4] S. Nadimi, B. Bhanu, "Physics-based models of color and IR video for sensor fusion", *Proc. IEEE Multisensor Fusion and Integration for Intelligent systems, MFI'03*, pp. 161-166, July 2003;

[5] I. A. Essa, "Ubiquitous sensing for smart and aware environments: technology towards the building of an aware home", *IEEE Personal Communications*, pp. 47-49, October 2000;

[6] J. W. Fisher III, T. Darrell, "Signal level fusion for multimodal perceptual user interface", *Proc. ACM PUI 2001*, Orlando, FL USA;

[7] A. Garg, V. Pavlovic, J. Rehg, "Boosted learning in dynamic Bayesian networks for multimodal speaker detection", *Proc. IEEE*, Vol. 91, No. 9, pp. 1355-1369, Sep. 2003;

[8] D. G. Stork, G. Wolff, E. Levine**,** "Neural network lipreading system for improved speech recognition", *Proc. Intl. Conf. on Neural Networks, IJCNN'92*, Vol. 2, pp. 289-295, 1992;

[9] R. Cutler, L. Davis, "Look who's talking: Speaker detection using video and audio correlation", *Proc. IEEE Intl. Conf. Multimedia and Expo. ICME'00*, pp. 1589-1592, 2000;

[10] T. Ikeda, H. Ishiguro, M. Asada, "Attention to clapping - a direct method for detecting sound source from video and audio", *Proc. of IEEE Intl. Conf. on Multisensor Fusion and Integration for Intelligent Systems, MFI'03*, pp. 26- 268, 2003;

[11] R.O. Duda, P.E. Hart, and David G. Stork, *Pattern Classification*, John Wiley & Sons, New York, 2001.