Tracking Humans using Multi-modal Fusion

Xiaotao Zou, Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, CA 92521
{xzou,bhanu}@vislab.ee.ucr.edu

Abstract

Human motion detection plays an important role in automated surveillance systems. However, it is challenging to detect non-rigid moving objects (e.g. human) robustly in a cluttered environment. In this paper, we compare two approaches for detecting walking humans using multi-modal measurements-video and audio sequences. The first approach is based on the Time-Delay Neural Network (TDNN), which fuses the audio and visual data at the feature level to detect the walking human. The second approach employs the Bayesian Network (BN) for jointly modeling the video and audio signals. Parameter estimation of the graphical models is executed using the Expectation-Maximization (EM) algorithm. And the location of the target is tracked by the Bayes inference. Experiments are performed in several indoor and outdoor scenarios: in the lab, more than one person walking, occlusion by bushes etc. The comparison of performance and efficiency of the two approaches are also presented.

1. Introduction

Automated Surveillance addresses real-time observation of people, vehicles and other moving objects within a complicated 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. 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 [3], which consist of a single camera or hundreds of cameras. To achieve the continuous monitoring, infrared (IR) cameras are used along with the optical cameras under low illuminations [12]. Besides these video or IR surveillance systems, there also exist "detection and tracking" systems based on non-imaging measurements. A project named "Smart Floor" [6] 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.

Although the video or IR sensors provide a detailed and friendly description about the environment, their volumes and costs restrict the use of them in the Wireless Sensor Network (WSN). On the contrary, the microphone turns out to be a qualified candidate to the surveillance WSN in several aspects: its compact size, low cost, small data volume for transmission, low power consumption, and easiness of integration in the chip. In most surveillance scenarios, a simple sensor network composed of several off-the-shelf cameras and dozens, even hundreds, of microphones may be appropriate to cover an area up to a few acres. Since the audio and video sequences cover overlapping areas, a sensor fusion mechanism should be developed to obtain a more accurate and efficient solution.

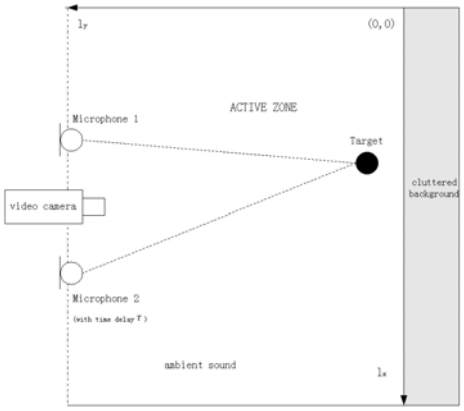

Figure 1. Experiment setup for the multi-modal surveillance system. l_x is the horizontal position of the target.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

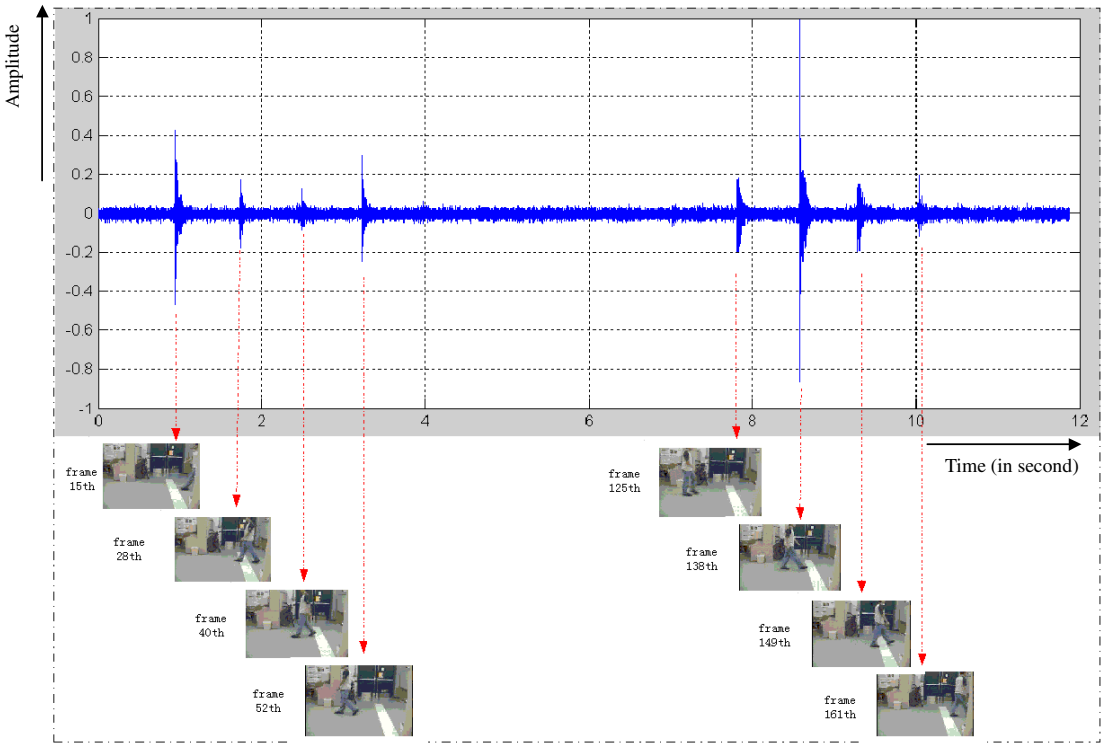

Figure 2. Audio-video data correspondence in the sequence "walking-in-the-lab". (Above) Sound pressure waveform received at the microphone. (Bottom) Corresponding frames for beats in the step sound.

In the multi-modal tracking system shown in Fig. 1, the audio waveforms are captured by two microphones, and the video sequences are recorded by the off-the-shelf camera. The frames contain a person walking in front of a cluttered background that may include other person. The audio waveform contains the object's step sounds corrupted by background noises.

The audio and visual signals are highly correlated as shown in Fig. 2. Also, the time delay between the signals arriving at the two microphones is correlated with the position of the walking person in the frames. In principle, tasks such as tracking may be performed better by taking advantage of these correlations. However, relevant features are not directly observable. The audio signal propagating from the walker is usually corrupted by reverberation, multi-path effects and background noise, making it difficult to measure the time delay. Moreover, the video sequence is cluttered by moving objects other than the walking person.

In this paper, we compare two multi-modal fusion approaches for walking human detection and tracking: Time-delay Neural Network (TDNN) and Bayesian Network (BN). First, we discuss related work and motivation of our approaches in Section 2. In the next section, the two approaches used for our problem are outlined. The network architecture, feature selection and parameter estimation are presented in details. Then, we compare the performance and efficiency of these two approaches on various testing scenarios: indoor, outdoors, more than one moving humans, occlusion by bushes.

2. Related work, motivation and contributions

2.1. Related work

Several studies on fusing audio-video data for object detection and tracking have been reported in the literature. Vermaak *et al.* [14] proposed the particle filter-based approach for audio-visual speaker tracking. Fisher and Darrell [7] presented an information theoretic approach for the fusion of multiple modalities, which can detect where a speaker is within a scene and whether he or she is producing specific words. A variation of the neural network, Time-Delay Neural Network (TDNN), has been proposed to handle the sequential data for multi-modal fusion. Stork *et al.* [13] proposed a modified TDNN to perform the visual lip-reading to improve the accuracy of acoustic speech recognition. Cutler and Davis [4] 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 objects.

Bayesian Network (BN) [10], the graphically statistical model, finds a wide use in multi-modal fusion. Garg *et al.* [9] developed a supervised learning framework based on dynamic Bayesian Networks and applied it to the problem of audio-visual speaker detection for a smart kiosk. A graphical model for audio-visual object tracking has been proposed by Beal and Jojic *et al.* [1], which extended the concept of Transformed Mixture of Gaussian (TMG) [8] to audio and video data modeling and used the Expectation-Maximization algorithm to estimate the model parameters.

2.2. Motivation

The visual motion of a walking person (i.e., gait) is periodic and highly correlated with the corresponding step sound (Fig. 2). A similar fact (correlation between the motion of mouths during speaking and the speech sounds) has been exploited for lip-reading [13] and speaker detection [4]. Fig. 3 shows the recurrence matrix of the extracted human motion and the similarity of the step sounds in the sequence "walking-in-the-lab". Recurrence matrix is a quantitative tool used to perform time series analysis of non-linear dynamic systems, and it is defined by the correlation function $R(I_{t_1}, I_{t_2})$ for frames I_{t_1}, I_{t_2} in our case. The highest correlation values can be found on the diagonal line in Fig. 3(a). Similarly, we use the Euclidean distance between the amplitudes to define the similarity of the corresponding audio signals at times t_1 and t_2. And the similarity is denoted by the brightness shown in Fig. 3(b). In Fig. 3, we can find 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, i.e. visual and audio.

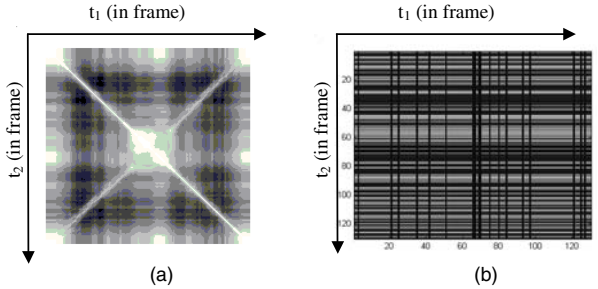

Figure 3. (a) Recurrence matrix $RM(t_1,t_2)$ of extracted object in video and (b) similarity of step sounds.

2.3. Contributions

This paper first explores the relation between visual motions and step sounds, and discusses the application of Time-Delay Neural Network (TDNN) in multi-modal fusion for walking human detection. The audio-visual correlation is first learned by a time-delay neural network, which then performs a spatio-temporal search over the audio-visual sequences for the walking person.

For comparison, the paper also employs the Bayesian Network for jointly modeling audio-visual data and exploiting correlations between the two modalities. Statistical models have several important advantages that make it suitable for our purpose. First, since we explicitly model the actual sources of variability in the problem, the resulting algorithm turns out to be robust. Second, using statistical models leads to an optimal solution by the Bayes inference. Third, parameter estimation is performed efficiently using the Expectation-Maximization algorithm.

3. Technical approaches

3.1. Time-Delay Neural Network approach

The Time-Delay Neural Network approach is illustrated in Fig. 4.

3.1.1. TDNN dataflow diagram

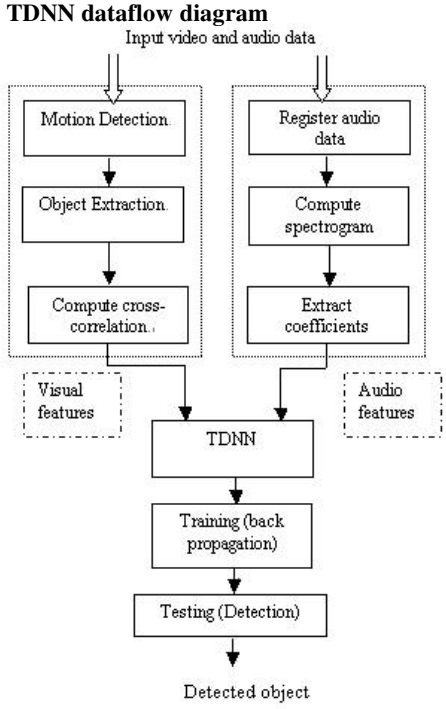

Figure 4. Dataflow diagram of TDNN-based object detection.

3.1.2. Time-Delay Neural Network architecture

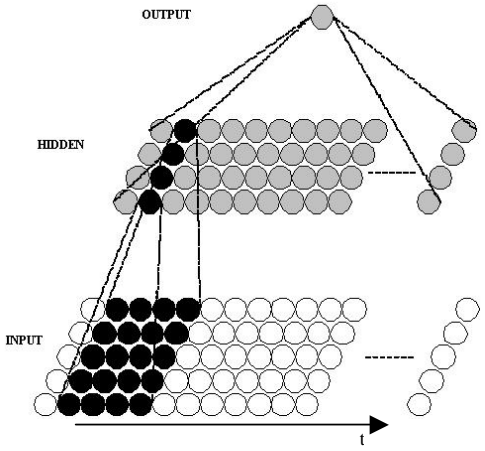

Figure 5. A typical Time-Delay Neural Network.

Fig. 5 shows a typical Time-Delay Neural Network 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).

3.1.3. Feature selection. As to visual input features for the TDNN, we choose a simple measure of change between two images I_t and I_{t+i} (the normalized cross-correlation):

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

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. All moving objects in the scene (walking or non-walking) have been first detected by the modified Background Subtraction algorithm [2]. By computing correlations between each subtracted object, we obtain the desired visual feature (frame cross-correlation $R_{t,t+1}$), which lies on the line immediately below the diagonal in Fig. 3(a).

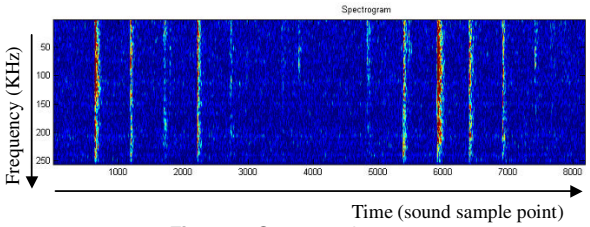

Figure 6. Step sound spectrogram.

The sound spectrogram (also called sonogram) coefficients of the step sounds are used as the audio input features. The sound spectrogram, like a musical score, is a visual representation of the sound. It is calculated by the short-time Fourier Transform (STFT) on the one-dimensional audio signal (shown in Fig. 6). Its horizontal dimension corresponds to time, and the vertical dimension denotes frequency. The relative intensity of the sound spectrogram at a particular time and frequency is indicated by the brightness at that point.

3.2. Bayesian Network approach

Bayesian Network (BN) is an attractive framework for statistical modeling, as it combines an intuitive graphical representation with efficient algorithms for inference and learning. BN encodes conditional dependences among a set of random variables in the form of a graph (e.g. Fig. 7). An arc between two nodes denotes a conditional dependence relationship, which is parameterized by a conditional probability model. The structure of the graph encodes domain knowledge, such as relationship between sensor outputs and hidden states, while the parameters of the conditional probability models can be learned from data. Another advantage of the BN models is that it can be easily extended to handle time series data, by means of the dynamic Bayesian Network (DBN) framework.

3.2.1. Video component. Video frames are modeled using a statistical model called Transformed Mixture of Gaussians (TMG) [8] (Fig. 7). This is a simple generative model that describes the observed image y in terms of an original image v that has been shifted by l_x pixels, and further contaminated by additive noise with covariance matrix Ψ. To account for the variability in the original image, v is modeled by a mixture model with components s. Component s consists of a template with mean μ_s and covariance matrix ϕ_s, and s has a prior probability π_s.

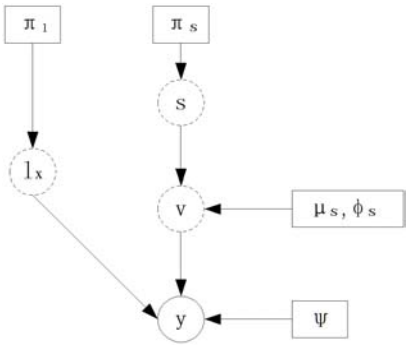

Figure 7. Graphical model of video signals.

And the parent-child relationships of the nodes are described by the conditional probabilities:

$$\Pr(v \mid s) = N(v \mid \mu_s, \phi_s), \Pr(s) = \pi_s,$$
$$\Pr(y \mid v, l_x) = N(y \mid G_{l_x} v, \Psi) \quad (2)$$

Here we assume both conditional probabilities $\Pr(v \mid s), \Pr(y \mid v, l_x)$ are Gaussians, in which $N(v \mid \mu_s, \phi_s)$ means a Gaussian distribution of variable v with the mean μ_s and covariance matrix ϕ_s. G_{lx} denotes the horizontal shift operator. And the prior probability for the shift l_x is assumed flat: $\Pr(l_x)=\pi_l$ (constant). The model parameters, including the image template μ_s, their covariance matrix ϕ_s, and the noise covariance Ψ, are learned from sequential data using the EM algorithm.

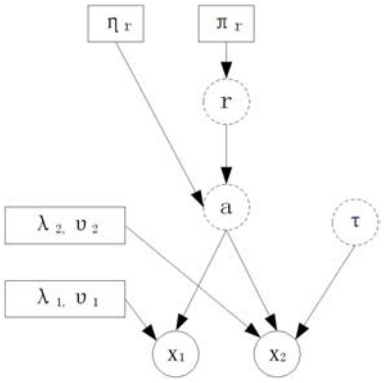

Figure 8. Graphical model of audio signals received by two microphones.

3.2.2. Audio component. Similarly, the audio model describes the observed audio signals x_1, x_2 in terms of an original signal a, which has been attenuated by a factor λ_i on its way to the microphone 1 and 2. It arrives at the microphone 2 with a time delay τ relative to that at microphone 1. To account for variability in the original signal, a is modeled by a mixture model with components r. Each component r has zero mean, covariance matrix η_r, and the prior probability π_r. Then we have:

$$\Pr(a \mid r) = N(a \mid 0, \eta_r), \Pr(r) = \pi_r,$$
$$\Pr(x_1 \mid a) = N(x_1 \mid \lambda_1 a, \upsilon_1), \quad (3)$$
$$\Pr(x_2 \mid a, \tau) = N(x_2 \mid \lambda_2 L_\tau a, \upsilon_2).$$

In which all conditional probabilities are assumed Gaussian, and L_τ denotes the temporal shift operator over original signal a.

3.2.3. Link between audio and video signals. The dependence of the time delay τ on the object location l_x in frames is modeled by a noisy linear mapping:

$$\Pr(\tau \mid l_x) = N(\tau \mid \alpha l_x + \beta, \upsilon_\tau) \quad (4)$$

In our experiment, the mapping involves only the horizontal position, as the vertical movement of the object has a significantly smaller effect on the arrival time compared to the horizontal motion. It can be shown that the linear approximation is fairly accurate for the pinhole camera and the large microphone baseline. To account for deviation from linearity and other inaccuracies in the simplified model, such as reverberation, we allow the mapping to be noisy, with a noise covariance matrix υ_τ. The audio-visual generative model is illustrated in Fig. 9.

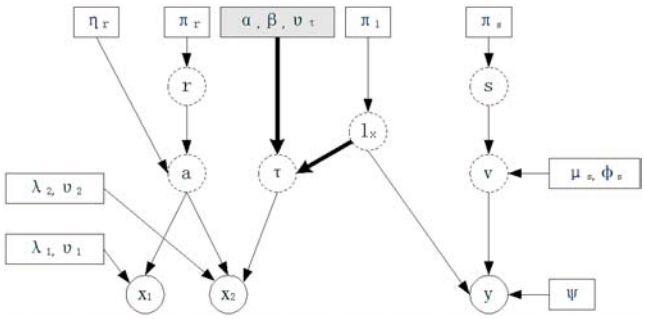

Figure 9. Graphical model representation of the full Bayesian Network to model both audio and video signals jointly. The arrows and parameters of the audio-visual link are highlighted.

To handle the sequential audio-visual data, we extend the static BN (Fig. 9) to the Dynamic BN, which obeys the property of independent identical distribution,

which means the model parameters are shared between different times.

3.2.4. Parameter estimation and Bayes inference

In the graphic model described above, the joint distribution of the observed signals, the unobserved variables and the component parameters, is given by:

$$\begin{aligned}&\Pr(x_1, x_2, y, \tau, l_x, r, s, a, v)\\ &= \Pr(x_1 \mid a)\Pr(x_2 \mid a, \tau)\Pr(a \mid r)\Pr(r) \quad \text{(Audio signal)}\\ &\quad * \Pr(y \mid v, l_x)\Pr(v \mid s)\Pr(s) \quad \text{(Video signal)}\\ &\quad * \Pr(\tau \mid l_x)\Pr(l_x) \quad \text{(Audio-visual link)}\end{aligned} \tag{5}$$

which is the product of the distributions defined by the audio and the video models. The model parameters $\Theta = \{\lambda_1, \upsilon_1, \lambda_2, \upsilon_2, \eta_r, \pi_r, \pi_s, \pi_l, \mu_s, \phi_s, \Psi, \alpha, \beta, \upsilon_\tau\}$ are estimated from the data sequence using the Expectation-Maximization (EM) algorithm [1].

After estimating the parameters, the *a posteriori* probability $\Pr(l_x \mid x_1, x_2, y)$ of the location variable l_x is calculated using the Bayes' rule:

$$\Pr(l_x \mid x_1, x_2, y) = \frac{\Pr(l_x, x_1, x_2, y \mid \Theta)}{\Pr(x_1, x_2, y \mid \Theta)} \tag{6}$$

The estimate of the object's location at each frame is the most likely estimate given the observed data:

$$\tilde{l}_x = \arg\max_{l_x} \Pr(l_x \mid x_1, x_2, y) \tag{7}$$

4. Experimental results

4.1. Experiment setup and devices

The video sequences are recorded using a SONY® DCR-VX1000 Digital Handycam camcorder. It has three 1/3-inch CCDs with 410,000 pixels for each. The resolution of frames is 720*480 pixels. The video capture rate is 30 frames per second. The sound was recorded with the built-in microphone of the DCR-VX1000 camcorder. A LABTEC® VERSE 303 microphone is also used as an auxiliary audio recording device when necessary. The sampling rate of both microphones is 32KHz at the resolution of 12-bit. And the video camera and microphones are mounted on the fixed platform.

The Time-Delay Neural Network has been implemented with the Neural Network Toolbox in MATLAB 6.1. And the Bayesian Network in this paper uses the Bayes Net Toolbox for MATLAB [11]. The testing platform is a Pentium IV 1.7 GHz PC with 256 MB (PC2100) memory.

4.2. Experiment scenarios

To test the performance of our approaches, we designed several scenarios for the experiment (description listed in Table 1 and shown in Fig. 11). The basic assumption in our experiments is that there is only one human (the "object") walking in the indoor or outdoor environment, which is monitored by an off-the-shelf video camera and several microphones. The environment could be complicated and cluttered, in which there may be other humans moving or walking. Our ultimate goal is to detect and track the object that is moving and producing step sound which is recorded.

Table 1: Description of test scenarios

Test scenarios name	Indoor/ outdoor	Humans in scene	Test sequences size (frames)
"Lab"	Indoor	2	160
"Two-person-test1"	Outdoor	2	120
"Two-person-test2"	Outdoor	2	260
"Behind-bush-test1"	Outdoor	1	195
"Behind-bush-test2"	Outdoor	1	229

To train the TDNN and BN, training sequences are approximately three times the size of test sequences. In TDNN, we choose one visual feature ($R_{t,t+1}$) and four audio features (spectrogram coefficients at 1, 1K, 10K and 100K Hz) as input features. And as the TDNN input specification, the audio features are normalized to the range [0 1]. After obtaining both visual and audio features from the training sequence, we fed them into the TDNN described earlier to train the weights $\{W_1\ W_2\ \ldots\ W_{84}\}$. Then, the spatio-temporal search with the learned TDNN is performed in each frame for locating the walking human. Fig. 11(a) shows the location of the detected target (marked with cross and rectangle) in the test sequence "Lab".

In our Bayesian Network approach, since there is only one object in the scene, we assume a single visual template *s* and a single audio component *r*. The EM algorithm is used to train the graphical model (Fig. 9) based the observed multi-modal signals. The maximum number of iterations is set to be 10, and the iterative operations are terminated when a small stopping criterion (10^{-10}) is met. If the estimated target location $\tilde{l}_x$ is within 30 pixels of the ground truth l_x, we consider it the correct detection of walking human. The ground truth, tracking results and the confidence plot of the test sequence "Two-person-test2" are shown in Fig. 10.

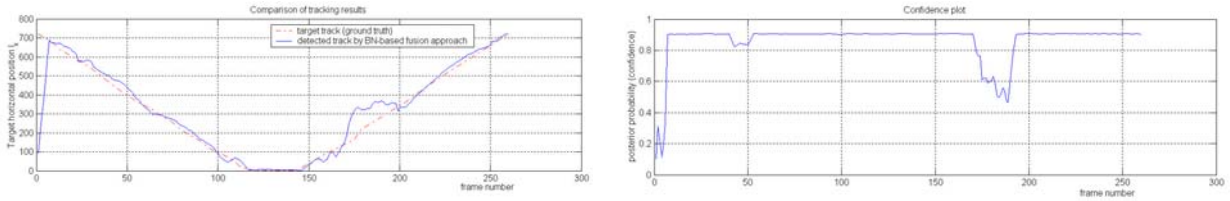

Figure 10. Tracking results of test sequence "Two-person-test2". Comparison of detected track and ground truth (left) and the confidence (*a posteriori* probability) plot (right).

(a) "Lab" (tracking results are marked in each frame)

(b) "Two-person-test1"

(c) "Two-person-test2"

(d) "Behind-bush-test1"

(e) "Behind-bush-test2"

Figure 11. Sample frames of all test sequences. The tracked object centers are highlighted as white blobs in each frame.

4.3. Performance comparison of TDNN and BN approaches

Table 2: Comparison of detection accuracy rate

Test scenarios	TDNN approach	BN approach
"Lab"	89%	93%
"Two-person-test1"	91%	95%
"Two-person-test2"	52%	86%
"Behind-bush-test1"	48%	83%
"Behind-bush-test2"	39%	72%

Table 3: Efficiency (training time) comparison

Test scenarios	TDNN (in second)	BN (in second)
"Lab"	327	348
"Two-person-test1"	228	276
"Two-person-test2"	600	528
"Behind-bush-test1"	453	411
"Behind-bush-test2"	508	472

5. Conclusion

In this paper, we discuss the use of multi-modal fusion for human motion detection. Specifically, we present two approaches for detecting walking humans in the cluttered environment based on video sequences and step sounds: the Time-Delay Neural Network (TDNN) and Bayesian Network (BN) approaches.

The comparison of these two approaches illustrates the advantages of statistical models (i.e. Bayesian Network) over the Time-Delay Neural Network: First, it's necessary to initialize the object in the video signals (i.e. pre-detection) in TDNN approach. In contrast, there is only random parameter initialization in the BN approach, and the choosing of the initial parameters doesn't affect the performance of the BN approach. Secondly, in the BN approach, there is a microphone array (Microphone 1 and 2), and the correlation between the time delay in audio signals and the object position in video sequences is modeled with a noisy linear mapping. The Bayesian Network encodes this property in its structure and uses the Transformed Mixture of Gaussians to model both video and audio data. Thirdly, The explicit and easily accessible structure of graphical models is clearly an advantage, while the inner structure and parameters of the TDNN is not directly available to designers. Finally, besides the better tracking accuracy of the BN approach in all experimental scenarios, there is the confidence (*a posteriori* probability of the estimates) as the quantitative measure of the support to the decision.

The work presented in this paper can be extended in several ways. The microphone array could be employed in the modified TDNN approach. And the periodicity of the step sounds could be detected with the power spectrum estimation techniques for the detection of walking humans in audio sequences. Audio signal separation and processing techniques can be included for the multiple object tracking. Furthermore, when heterogeneous sound sources (e.g., vehicles) are also present, we may include seismic sensors in our scheme.

6. References

[1] M. Beal, N. Jojic, H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 25, No. 7, pp. 828- 836, July 2003.

[2] B. Bhanu and X. Zou, "Moving humans detection based on multi-modal sensory fusion," *Proc. IEEE Workshop on Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS'04)*, pp. 101- 108, July 2004.

[3] R. T. Collins, A. J. Lipton, H. Fujiyoshi and T. Kanade, "Algorithms for cooperative multisensor surveillance," *Proceedings of the IEEE*, Vol. 89, No. 10, pp. 1456-1477, October 2001.

[4] 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.

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

[6] 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.

[7] J. W. Fisher III, T. Darrell, "Signal level fusion for multimodal perceptual user interface," *Proc. Workshop on Perceptive User Interfaces* (*PUI '01)*, Nov. 2001.

[8] B. J. Frey, N. Jojic, "Transformation-invariant clustering using the EM algorithm," *IEEE Tran. Pattern Analysis and Machine Intelligence*, Vol. 25, No. 1, pp. 1- 17, January 2003.

[9] A. Garg, V. Pavlovic and J. Rehg, "Boosted learning in Dynamic Bayesian Networks for multimodal speaker detection," *Proceedings of the IEEE,* Vol. 91, No. 9, pp. 1355- 1369, September 2003.

[10] F. V. Jensen, *An Introduction to Bayesian Networks*, Springer-Verlag, New York, 1996.

[11]1K. Murphy, *Bayes Net Toolbox for MATLAB*, www.ai.mit.edu/~murphyk/Software/BNT/bnt.html

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

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

[14] J. Vermaak, M. Gangnet, A. Blake, P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," *Proc. IEEE Intl. Conf. Computer Vision (ICCV'01)*, Vol. 1, pp. 741- 746, July 2001.