Information Assimilation Framework for Event Detection in Multimedia Surveillance Systems

Pradeep Kumar Atrey
School of Computing
National University of Singapore
Republic of Singapore
pkatrey@nus.edu.sg

Mohan S. Kankanhalli
School of Computing
National University of Singapore
Republic of Singapore
mohan@comp.nus.edu.sg

Ramesh Jain
School of Information and Computer Sciences
University of California
Irvine, CA, USA
jain@ics.uci.edu

ABSTRACT

Most multimedia surveillance and monitoring systems nowadays utilize multiple types of sensors to detect events of interest as and when they occur in the environment. However, due to the asynchrony among and diversity of sensors, information assimilation - how to combine the information obtained from asynchronous and multifarious sources is an important and challenging research problem. In this paper, we propose a framework for information assimilation that addresses the issues - "when", "what" and "how" to assimilate the information obtained from different media sources in order to detect events in multimedia surveillance systems. The proposed framework adopts a hierarchical probabilistic assimilation approach to detect atomic and compound events. To detect an event, our framework uses not only the media streams available at the current instant but it also utilizes their two important properties - first, accumulated past history of whether they have been providing concurring or contradictory evidences, and - second, the system designer's confidence in them. The experimental results show the utility of the proposed framework.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]

General Terms

Security

Keywords

Information assimilation, Multimedia surveillance, Agreement coefficient, Confidence fusion, Event detection, Compound and atomic events

1. INTRODUCTION

In recent times, it is being increasingly accepted that most surveillance and monitoring tasks can be better performed by using multiple types of sensors as compared to using only a single type. Therefore, most surveillance systems nowadays utilize multiple types of sensors like microphones, motion detectors and RFIDs etc in addition to the video cameras. However, different sensors usually provide the sensed data in different formats and at different rates. For example, a video may be captured at a frame rate which could be different from the rate at which audio samples are obtained, or even two video sources can have different frames rates. Moreover, the processing time of different types of data is also different. Due to the asynchrony and diversity among streams, the assimilation of information in order to accomplish an analysis task is a challenging research problem. *Information assimilation refers to a process of combining the sensory and non-sensory information obtained from asynchronous multifarious sources using the context and past experience.*

Event detection is one of the fundamental analysis tasks in multimedia surveillance and monitoring systems. In this paper, we propose an information assimilation framework for event detection in multimedia surveillance and monitoring systems.

Events are usually not impulse phenomena in real world, but they occur over an interval of time. Based on different granularity levels in time, location, number of objects and their activities, an event can be a "compound-event" or simply an "atomic-event". We define compound-events and the atomic-events as follows -

Definition 1. Event is a physical reality that consists of one or more living or non-living real world objects (who) having one or more attributes (of type) being involved in one or more activities (what) at a location (where) over a period of time (when).

Definition 2. Atomic-event is an event in which exactly one object having one or more attributes is involved in exactly one activity.

Definition 3. Compound-event is the composition of two or more different atomic-events.

A compound-event, for example, "a person is running and

shouting in the corridor" can be decomposed into its constituent atomic-events - "a person is running in the corridor" and "a person is shouting in the corridor". The atomic-events in a compound event can occur simultaneously, as in the example give above; or they may also occur one after another, for example, the compound-event "A person walked through the corridor, stood near the meeting room, and then ran to the other side of the corridor" consists of three atomic-events "a person walked through the corridor" followed by "person stood near the meeting room", and then followed by "person ran to the other side of the corridor".

The different atomic-events, to be detected, may require different types of sensors. For example, a "walking" and "running" event can be detected based on video and audio streams, a "standing" event can be detected using video but not by using audio streams, and "shouting" event can be better detected using the audio streams. The different atomic-events require different minimum time-periods over which they can be confirmed. This minimum time-period for different atomic-events depends upon the time in which the amount of data sufficient to reliably detect an event can be obtained and processed. Even the same atomic-event can be confirmed in different time periods using different data streams. For example, minimum video data required to detect a walking event could be of two seconds; however, the same event can be detected based on audio data of one second.

The media streams in a multimedia system are often correlated. We assume that the system designer has a confidence level in the decision obtained based on each of the media streams; and there is a cost of obtaining these decisions which usually includes the cost of sensor, its installation and maintenance cost, the cost of energy to operate it, and the processing cost of the stream. We also assume that each stream in a multimedia system partially helps in accomplishing the analysis task (e.g. event detection). The various research issues in the assimilation of information in such systems are -

1. *When to assimilate?.* Events occur over a timeline [Chieu and Lee 2004]. Timeline refers to a measurable span of time with information denoted at designated points. Timeline-based event detection in multimedia surveillance sys- tems requires identification of the designated points along a timeline at which assimilation of information should take place. Identification of these designated points is challenging because of asynchrony and diversity among streams and also because of the fact that different events have different granularity levels in time.

2. *What to assimilate?* The fact that, at any instant all of the employed media streams do not necessarily contribute towards accomplishing the analysis task brings up the issue of finding the most informative subset of streams. From the available set of streams,

   - What is the optimal number of streams required to detect an event under the specified constraints?
   - Which subset of the streams is the optimal one?
   - In case the most suitable subset is unavailable, can one use alternate streams without much loss of cost-effectiveness and confidence?
   - How frequently should this optimal subset be computed so that the overall cost of the system is minimized?

3. *How to assimilate?* In combining of different data sources,

   - How to utilize the correlation among streams?
   - How to integrate the contextual information (such as environment information) and the past experience?

The framework for information assimilation, which we propose, essentially addresses the above-mentioned issues. Note that, the solution to the issue (2) has been described with detailed results and analysis in our other work [Atrey et al. 2006]. In this paper, we focus on issues (1) and (3) and present our framework for information assimilation with detailed analysis and results[1].

The proposed framework for information assimilation has the following distinct characteristics -

- The detection of events based on individual streams are usually not accomplished with certainty. To obtain a binary decision, early thresholding of uncertain information about an event may lead to error. For example, let an event detector find the probabilities of the occurrence of an event based on three media streams M_1, M_2 and M_3, to be 0.60, 0.62 and 0.70, respectively. If the threshold is 0.65, then these probabilistic decisions are converted into binary decisions 0, 0 and 1, respectively; which implies that the event is found occurring based on stream M_3 but is found non-occurring based on stream M_1 and M_2. Since two decisions are in favor of non-occurrence of event compared to the one decision in favor of occurrence of the event, by adopting a simple voting strategy, the overall decision would be that the event did not occur. It is important to note that early thresholding can introduce errors in the overall decision. In contrast to early thresholding, the proposed framework advocates late thresholding by first assimilating the probabilistic decisions that are obtained based on individual streams, and then by thresholding the overall probability (which is usually more than the individual probabilities e.g. 0.85 in this case) of occurrence of event based on all the streams, which is less erroneous.
- The sensors capturing the same environment usually provide concurring or contradictory evidences about what is happening in the environment. The proposed framework utilizes this agreement/disagreement information among the media streams to strengthen the overall decision about the events happening in the environment. For example, if two sensors have been providing concurring evidences in the past, it makes sense

[1]The earlier version of some of the results found here was published in [Atrey et al. 2005]

to give more weight to their current combined evidence compared to the case if they provided contradictory evidences in the past [Siegel and Wu 2004]. The agreement/disagreement information (we call it as "agreement coefficient") among media streams is computed based on how they have been agreeing or disagreeing in their decisions in the past. We also propose a method for fusing the agreement coefficients among the media streams.

- The designer of a multimedia analysis system can have different confidence levels in different media streams for accomplishing different tasks. The proposed framework utilizes the confidence information by assigning a higher weight to the media stream which has a higher confidence level. The confidence in each stream is computed based on how accurate it has been in the past. Integrating confidence information in the assimilation process also requires the computation of the overall confidence in a group of streams, a method for which is also proposed.

- Information assimilation is different from information fusion in that the former brings the notion of integrating context and the past experience in the fusion process. The context is an accessory information that helps in the correct interpretation of the observed data. We use the geometry of the monitored space along with the location, orientation and coverage area of the employed sensors as the spatial contextual information. We integrate the past experience by modelling the agreement/disagreement information among the media streams based on the accumulated past history of their agreement or disagreement.

Our contributions in this paper are as follows. We have identified various research issues which are important and challenging in assimilating the information for event detection in multimedia surveillance systems, and proposed a framework that adopts a hierarchical probabilistic approach to address these issues. The proposed framework has introduced the notion of compound and atomic events that helps in describing events over a timeline. Our probabilistic framework has not only utilized the agreement/disagreement information among the media streams, but it has also integrates their confidence information in the assimilation process, which helps in improving the overall accuracy of event detection. We have formulated the computation and fusion of the agreement coefficients among the streams and have also proposed a method for confidence fusion.

Rest of this paper is organized as follows. In section 2, we discuss the related work. We present our framework in section 3. The experimental results are reported in section 4. Finally, we conclude the paper with a discussion on future work in section 5.

2. RELATED WORK

Researchers have used early fusion as well as late fusion strategies in solving diverse problems. For example, feature-level (early) fusion of video and audio has been proposed for the problems speech processing [Hershey et al. 2004] and recognition [Nefian et al. 2002], tracking [Checka et al. 2004], and monologue detection [Nock et al. 2002] by using the mutual information among the video and audio features under the assumption that audio and video signals are individually and jointly Gaussian random variables. On the other hand, late fusion strategies have also been used in sensor fusion applications [Rao and Whyte 1993], [Chair and Varshney 1986], [Kam et al. 1992]. In late fusion strategy, a global decision is made by fusing the local decisions obtained from each data source. [Rao and Whyte 1993] presented a sensor fusion algorithm for identification of tracked targets in a decentralized environment. [Chair and Varshney 1986] established an optimal fusion rule with the assumption that each local sensor made a predetermined decision and each observation was independent. [Kam et al. 1992] generalizes their solution for fusing the correlated local decisions.

Similar to [Wu et al. 2004], we employ early (feature level) assimilation as well as late (decision level) assimilation strategy. Since each media stream provides various features (such as blob's location and area in case of a video stream), their assimilation is performed locally for each media stream to obtain a local decision. Once all the local decisions are available, a global decision is derived by assimilating the local decisions incorporating their agreement and confidence information. *The late assimilation strategy has an advantage over early assimilation in that the former offers scalability (i.e. graceful upgradation or degradation) in terms of media streams used in the assimilation process* [Atrey et al. 2006]. Note that, in late assimilation, we consider the media streams to be "decision-wise correlated". The decision-wise correlation refers to how the decisions obtained based on different media streams co-vary with each other.

Our work is different from the works cited above in following aspects. We explicitly compute and utilize the correlation information (we call it the "agreement coefficient") among the streams. Agreement coefficient among streams is computed based on how concurring or contradictory evidences they provide. Intuitively, higher the agreement among the streams, more would be the confidence in the global decision, and vice versa [Siegel and Wu 2004]. The various forms of correlation coefficients that have been used for diverse applications are based on content-wise dependency between the sources, hence are not suitable in our case. Pearson's correlation coefficient, Lin's concordance correlation coefficient [Lin 1989] and Kappa coefficient [Bloch and Kraemer 1989] cannot be used in our case since they are evaluated to zero when the covariance among the observations is zero. Therefore, the proposed framework models the agreement coefficient and its evolution based on the accumulated past history of how agreeing or disagreeing the media streams have been in their decisions.

Also, the past works in multimodal fusion literature do not consider the notion of having confidences in the different modalities. We incorporate the stream's confidence information. Recently, [Siegel and Wu 2004] has also pointed out the importance of considering the confidence in sensor fusion. The authors have used the Dempster-Shafer (D-S) 'theory of evidence' to fuse the confidences. In contrast, we propose a model for confidence fusion by using a Bayesian formulation because it is both simple and computationally efficient [Rao and Whyte 1993].

3. PROPOSED FRAMEWORK

3.1 Overview

The proposed information assimilation framework adopts a hierarchical probabilistic approach in order to detect an event in a surveillance and monitoring environment, and performs assimilation of information at three different hierarchical levels - media-stream level, atomic-event level and the compound-event level. The work flow of the framework is depicted in figure 1. Let a surveillance and monitoring system consists of n heterogeneous sensors that capture data from the environment. We employ n *Media Stream Processors* (MSP_1 to MSP_n), where each MSP_i, $1 \leq i \leq n$, is a set of media processing tools that extracts features from the media stream M_i; for example, a blob detector extracts blobs from a video stream. The features extracted from each media stream are stored in their respective databases.

Let the system detect N_a number of atomic-events. The total number of sets containing two or more atomic events in which the atomic events can occur together can be given by $\sum_{r=2}^{N_a} \binom{N_a}{r}$. Any k^{th} compound event $\mathbf{E}_k$ can be expressed as $\mathbf{E}_k = \langle \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_r \rangle$, where $2 \leq r \leq N_a$, $1 \leq k \leq N_c$, N_c being the number of compound events which can be detected by using the system. The total number N_E of events (atomic events as well as compound events) can be given by $N_E = N_a + N_c$.

A compound-event $\mathbf{E}_k$, which comprises of two or more atomic-events occurring together, is detected hierarchically in a bottom-up manner. First, atomic-events $\mathbf{e}_j$, $1 \leq j \leq r$ are detected using the relevant media streams, and then these decisions are assimilated hierarchically to obtain an overall decision for the compound event $\mathbf{E}_k$, as described in the subsequent subsections.

From the total number N_a of atomic events that the system can detect, the proposed framework identifies -

- The atomic events (e.g. person's standing/walking/ running and person's talking/shouting) that cannot occur simultaneously.
- The atomic events (e.g. person's walking) that can occur individually as well as can occur together with some other atomic event (e.g. with person's shouting).
- The atomic events (such as person's shouting) that cannot occur individually and must occur together with some other atomic event (such as with person's standing/walking/running).

To further illustrate it, we provide the following example.
Example 1: Let us consider a surveillance system that uses two types of sensors - video and audio with the goal of detecting $N_a = 6$ atomic events, namely - person's "standing", "walking", "running", "talking", "shouting" and "door knocking". In this case, as shown in Table 1, there could be $N_c = 9$ compound events in which any $r \geq 2$ atomic event(s) could occur. In total, there could be $N_E = 12$ events. Next, we also identify the types of data sources which can be used to detect each of the atomic events. For instance, the atomic events shown in example 1, can be detected as follows - standing (V), walking (AV), running (AV), talking (A), shouting (A), door knocking (A); where (A), (V) and (AV) denote audio, video and audio-video streams, respectively.

Table 1: All possible events in Example 1

Event number	Constituent atomic events
1	Standing
2	Walking
3	Running
4	Standing , Talking
5	Standing, Shouting
6	Standing, Door knocking
7	Walking, Talking
8	Running, Talking
9	Walking, Shouting
10	Running, Shouting
11	Standing, Talking, Door knocking
12	Standing, Shouting, Door knocking

3.2 Timeline-based event detection

As discussed earlier, the events occur over a timeline. There are various issues related to timeline-based event detection such as -

- To mark the start and end of an event over a timeline, there is a need to obtain and process the data streams at certain time intervals. This time interval, which is basically the minimum amount of time to confirm an event, could be different for different atomic/compound events when detected using different data streams. Determining the minimum time period (say t_w) to confirm different events is a research issue which is out of scope of this paper and will be explored in the future work. In this paper, we assume this minimum time period t_w to be the same for all the atomic/compound events.
- Determining the minimum time period t_w for a specific atomic event is also critical. Ideally, t_w should be just large enough to capture the data to confirm an event, since a small value of t_w allows to detect the events at a finer granularity in time. We learn the suitable value of t_w through experiments.
- Since the information from different sources become available at different time instances, when should it be assimilated is another research issue. There could several strategies to resolve this issue. We assimilate the information at fixed time intervals t_w. This time interval is determined by choosing the maximum of all the minimum time periods in which various atomic events can be confirmed. Although this strategy may not be the best, but is computationally less-expensive. Again, exploring other strategies is an issue which will be considered in the future.

3.3 Hierarchical probabilistic assimilation

The proposed framework adopts a hierarchical probabilistic assimilation approach and performs assimilation of information obtained from diverse data sources at three different levels - Media-stream level, Atomic-event level and Compound-event level. The details are as follows.

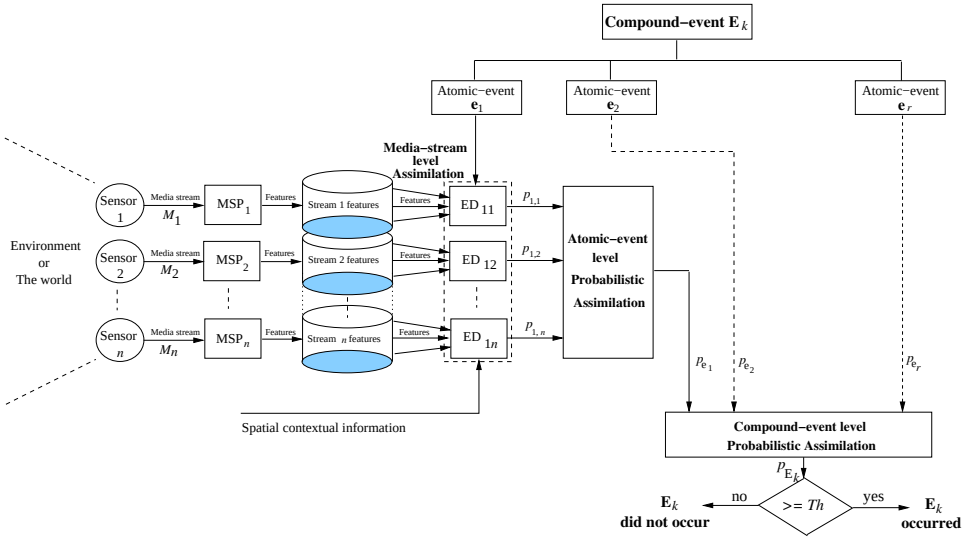


Figure 1: A schematic overview of the information assimilation framework for event detection

3.3.1 Media-stream level assimilation

The *Event Detectors* (ED_{ji}, $1 \leq j \leq r$ and $1 \leq i \leq n$) are employed to independently detect each atomic-event $\mathbf{e}_j$ based on the respective features obtained from media streams M_i, $1 \leq i \leq n$. At media-stream level, all the available features from a media stream are combined. The event detectors make the decision about an atomic event based on the combined features. Whenever required, they also utilize the environment information such as the geometry of the monitored space, location, orientation and the coverage space etc of sensors. The event detectors provide their decisions in probabilities $p_{j,i}$, $1 \leq j \leq r$ and $1 \leq i \leq n$ (Figure 1). The $p_{j,i}$ implies probability of the occurrence of atomic-event $\mathbf{e}_j$ based on media stream M_i.

3.3.2 Atomic-event level assimilation

At the next level, since the decisions about an atomic-event $\mathbf{e}_j$, that are obtained based on all the relevant media streams, may be similar or contradictory; these decisions are assimilated using a Bayesian approach incorporating streams' agreement/disagreement and confidence information. For the atomic-events $\mathbf{e}_j$, $1 \leq j \leq r$, we follow the steps -

1. At any particular instant, we group all the streams into two subsets S_1 and S_2. S_1 and S_2 contain the streams based on which the event detectors provide decision in favor and against the occurrence of the atomic-event, respectively.

2. Using the streams in the two subsets S_1 and S_2, we compute overall probabilities $P(\mathbf{e}_j|S_1)$ and $P(\bar{\mathbf{e}_j}|S_2)$ of occurrence and non-occurrence of the atomic-event $\mathbf{e}_j$, respectively. The overall probabilities are computed using a Bayesian assimilation approach which will be described shortly.

3. If $P(\mathbf{e}_j|S_1) \geq P(\bar{\mathbf{e}_j}|S_2)$, it is concluded that the atomic-event $\mathbf{e}_j$ has occurred with a probability $P_{\mathbf{e}_j} = P(\mathbf{e}_j|S_1)$, else it did not occur with a probability $P_{\mathbf{e}_j} = P(\bar{\mathbf{e}_j}|S_2)$.

We assume the media streams to be "content-wise" independent. This assumption is reasonable since media streams may be of different types, and may have different data formats and representations. However, since the decision about the same atomic-event is obtained based on all the streams, we can assume them to be "decision-wise" correlated.

We describe in the following paragraphs how the assimilation of decision-wise correlated media streams takes place, and also how the agreement coefficient and confidence information about them are modelled.

A. Assimilation of correlated media streams

Let a surveillance and monitoring system utilize a set $\mathbf{M}^n = \{M_1, M_2, \ldots, M_n\}$ of n media streams. The system outputs local decisions $P(\mathbf{e}_j|M_i)$, $1 \leq i \leq n$, $1 \leq j \leq r$, about an atomic-event $\mathbf{e}_j$. Along a timeline, as these probabilistic decisions are available, we iteratively integrate all the media streams using a Bayesian approach. The proposed approach allows for incremental and iterative addition of new stream. Let $P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i-1})$ denote probability of the occurrence of atomic-event $\mathbf{e}_j$ at time t based on from media streams $M_1, M_2, \ldots, M_{i-1}$. The updated probability $P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i})$ (i.e. the overall probability after assimilating the new stream $M_{i,t}$ at time instant t) can be iteratively computed as -

$$P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i}) = \frac{P(M_{i,t}|\mathbf{e}_{j_t})P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i-1})}{P(M_{i,t}|\mathbf{M}_t^{i-1})}$$

$$P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i}) = \alpha_i P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i-1})P(M_{i,t}|\mathbf{e}_{j_t}) \quad (1)$$

where, α_i is a normalization factor.

Equation (1) shows the assimilation using the Bayesian approach under the assumption that all the media streams have equal confidence levels and zero agreement coefficient. In what follows, we relax this assumption and integrate the agreement /disagreement and confidence information of media streams in their assimilation.

The confidence in each media stream is computed by experimentally determining its accuracy. To integrate the confidence into assimilation process, we use the consensus theory. Consensus theory provides a notion of combining the single probability distributions based on their weights [Benediktsson and Kanellopoulos 1999]. In our case, we essentially do the same by assigning weights to different media streams based on their confidence information. If we have more confidence in a media stream, a higher weight is given to it. Several consensus rules have been proposed, however the most commonly used consensus rules are - *linear opinion pool*(LOP) and *logarithmic opinion pool* (LOGP). In linear opinion pool, non-negative weights are associated with the sources to quantitatively express the "goodness" of each source. The rule is formulated as: $T_c(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} w_i p_i$ where, $p_i, 1 \leq i \leq n$, are the individual probabilistic decisions; and w_i, $1 \leq i \leq n$ are their corresponding weights whose sum is equal to 1 i.e. $\sum_{i=1}^{n} w_i = 1$. We use the *logarithmic opinion pool* since it satisfies the assumption of conditional (content-wise) independence among media streams which is essential to assimilation. The rule is described as [Genest and Zidek 1986] -

$$log[T_c(p_1, p_2, \ldots, p_n)] = \sum_{i=1}^{n} w_i log(p_i) \tag{2}$$

or

$$T_c(p_1, p_2, \ldots, p_n) = \prod_{i=1}^{n} {p_i}^{w_i} \tag{3}$$

where, $p_i, 1 \leq i \leq n$, are the individual probabilistic decisions and $\sum_{i=1}^{n} w_i = 1$. We normalize it over the two aspects of an event - the occurrence and non-occurrence of event. The formulation is shown as -

$$T_c(p_1, p_2, \ldots, p_n) = \frac{\prod_{i=1}^{n} {p_i}^{w_i}}{\sum_E (\prod_{i=1}^{n} {p_i}^{w_i})} \tag{4}$$

We use this formulation to develop the assimilation model which will be described shortly.

The agreement coefficient between two media streams is used as a scaling factor for the overall probability of occurrence of an event. The idea is that higher the agreement coefficient between the two media streams, the higher would be the overall probability. We use this notion in the proposed assimilation model.

The assimilation model that combines the probabilistic decisions based on two sources $\mathbf{M}^{i-1}$ (i.e. a group of $i-1$ streams) and M_i (i.e. an individual i^{th} stream) is given as follows-

$$P_i = \frac{(P_{i-1})^{F_{i-1}}.(p_i)^{f_i}.e^{\overline{\gamma}_i}}{(P_{i-1})^{F_{i-1}}.(p_i)^{f_i}.e^{\overline{\gamma}_i} + (1-P_{i-1})^{F_{i-1}}(1-p_i)^{f_i}.e^{-\overline{\gamma}_i}} \tag{5}$$

where, $P_i = P(\mathbf{e}_{j_t}|\mathbf{M}_t^i)$ and $P_{i-1} = P(\mathbf{e}_{j_t}|\mathbf{M}_t^{i-1})$ are the probabilities of occurrence of atomic-event $\mathbf{e}_j$ using $\mathbf{M}^i$ and $\mathbf{M}^{i-1}$, respectively, at time instant t. $p_i = P(\mathbf{e}_{j_t}|M_{i,t})$ is probability of the occurrence of atomic-event $\mathbf{e}_j$ based on only i^{th} stream at time instant t. Similarly, F_{i-1} and f_i (such that $F_{i-1}+f_i = 1$) are the confidence in $\mathbf{M}^{i-1}$ and M_i, respectively. The computation of confidence for a group of media streams will be described shortly. The $\overline{\gamma}_i \in [-1, 1]$ is the agreement coefficient between two sources $\mathbf{M}^{i-1}$ and M_i. The limits -1 and 1 represent full disagreement and full agreement, respectively, between the two sources. The modelling of $\overline{\gamma}_i$ is described in subsequent paragraphs.

B. Modelling of the agreement coefficient

The correlation among the media streams refers to the measure of their agreement or disagreement with each other. We call this measure of agreement to be the "Agreement Coefficient" among the streams. Let the measure of agreement among the media streams at time t be represented by a set $\Gamma(t)$ which is expressed as:

$$\Gamma(t) = \{\gamma_{ik}(t)\} \tag{6}$$

where, the term $-1 \leq \gamma_{ik}(t) \leq 1$ is the *agreement coefficient* between the media streams M_i and M_k at time instant t.

The *agreement coefficient* $\gamma_{ik}(t)$ between the media streams M_i and M_k at time instant t is computed by iteratively averaging the past agreement coefficients with the current observation. The $\gamma_{ik}(t)$ is precisely computed as:

$$\gamma_{ik}(t) = \frac{1}{2}\left[(1 - 2 \times abs(p_i(t) - p_k(t))) + \gamma_{ik}(t-1)\right] \tag{7}$$

where, $p_i(t) = P(\mathbf{e}_{j_t}|M_i)$ and $p_k(t) = P(\mathbf{e}_{j_t}|M_k)$ are the individual probabilities of occurrence of atomic-event $\mathbf{e}_j$ based on media streams M_i and M_k, respectively, at time $t \geq 1$; and $\gamma_{ij}(0) = 1 - 2 \times abs(p_i(0) - p_k(0))$. These probabilities represent decisions about the atomic-events. Exactly same probabilities would imply full agreement ($\gamma_{ik} = 1$) whereas totally dissimilar probabilities would mean that the two streams fully contradict each other ($\gamma_{ik} = -1$). Note that any three media streams, in agreeing/disagreeing with each other, do follow the commutativity rule.

The agreement coefficient between two sources $\mathbf{M}^{i-1}$ and M_i is modelled as:

$$\overline{\gamma}_i = \frac{1}{i-1} \sum_{s=1}^{i-1} \gamma_{si} \tag{8}$$

where, γ_{si} for $1 \leq s \leq i-1$, $1 < i \leq n$ is the agreement coefficients between the s^{th} and i^{th} media streams. The agreement fusion model given in equation (8) is based on *average-link clustering*. In average-link clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster. In our case, a group $\mathbf{M}^{i-1}$ of $i-1$ media streams is one cluster and we find the average distance of new i^{th} media stream with this cluster. The fused agreement coefficient $\overline{\gamma}_i$ is used for combining M_i with $\mathbf{M}^{i-1}$ as described before in equation (5).

C. Confidence fusion

In the context of streams, the confidence in a stream is related to its accuracy. The higher the accuracy of a stream, higher the confidence we would have in it. We compute the accuracy of a stream by determining how many times an event is correctly detected based on it out of the total number of tries. Note that, in our case, the accuracy of a stream includes the measurement accuracy of the sensor as well as the accuracy of the algorithm used for processing the stream.

The *confidence fusion* refers to the process of finding the overall confidence in a group of media streams where the

individual media streams have their own confidence level. If the two streams M_i and M_k have their confidence levels f_i and f_k, respectively; what would our confidence be in a group which contains both the streams? The intuitive answer to this question would be that our overall confidence should increase as the number of streams increases. Considering the confidence values as the probabilities, we propose a Bayesian method to fuse the confidence levels in individual streams. The overall confidence f_{ik} in a group of two media streams M_i and M_k is computed as follows:

$$f_{ik} = \frac{f_i \times f_k}{f_i \times f_k + (1 - f_i) \times (1 - f_k)} \quad (9)$$

In the above formulation, we make two assumptions. First, we assume that the system designer's confidence level in each of the media streams is more than 0.5. This assumption is reasonable since there is no use of employing a sensor which is found to be inaccurate more than half of the time. Second, although the media streams are correlated in their decisions; we assume that they are mutually independent in terms of their confidence levels.

For n number of media streams, the overall confidence is iteratively computed. Let F_{i-1} be the overall confidence in a group of $i-1$ streams. By fusing the confidence f_i of i^{th} stream with F_{i-1}, the overall confidence F_i in a group of i streams is computed as:

$$F_i = \frac{F_{i-1} \times f_i}{F_{i-1} \times f_i + (1 - F_{i-1}) \times (1 - f_i)} \quad (10)$$

3.3.3 *Compound-event level assimilation*

At the compound-event level, the overall probability $p_{\mathbf{E}}$ of the occurrence of compound-event $\mathbf{E}$ is estimated by assimilating the probabilistic decisions p_{e_j}, $1 \leq j \leq r$ about the r atomic-events by using the following assimilation model -

$$p_{\mathbf{E}} = \frac{\prod_{j=1}^{r} p_{e_j}}{\prod_{j=1}^{r} p_{e_j} + \prod_{j=1}^{r} (1 - p_{e_j})} \quad (11)$$

If $p_{\mathbf{E}}$ is found greater than the threshold Th, the system decides in favor of the occurrence of compound event $\mathbf{E}$, else it decides against it.

Since the atomic-events are independent, the agreement coefficients among them are considered as zero, and hence is not integrated into equation (11). For example, atomic-events $\mathbf{e}_1$ = "A person is walking in the corridor" and $\mathbf{e}_2$ = "A person is shouting in the corridor" are essentially independent since a person's walking is completely independent of the person's shouting. The confidence information is also not integrated into this assimilation model because the confidence is usually associated with media streams and not with the atomic-events.

4. EXPERIMENTAL RESULTS

To demonstrate the utility of our proposed framework, we present experimental results in a surveillance and monitoring scenario. The surveillance environment is the corridor of our school building and the system goal is to detect events that are described in Example 1 (in section 3.1) i.e. human's running, walking, standing, talking, shouting and door knocking in the corridor. The environment layout is shown in figure 2. We use two video sensors (cameras M_1 and M_2) to record the video from the two opposite sides of corridor, and two audio sensors (microphones M_3 and M_4) to capture the ambient sound. A snapshot of the multimedia surveillance system which we have developed is shown in figure 3. The system is implemented using Visual C++ on MS-Windows platform. The MS-Access is used a database to store the features and the events.

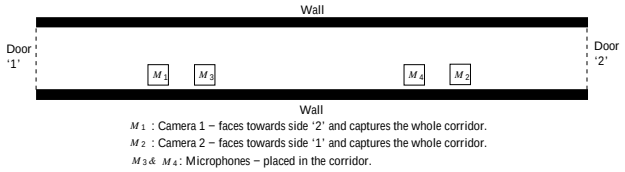


Figure 2: The layout of the corridor under surveillance and monitoring

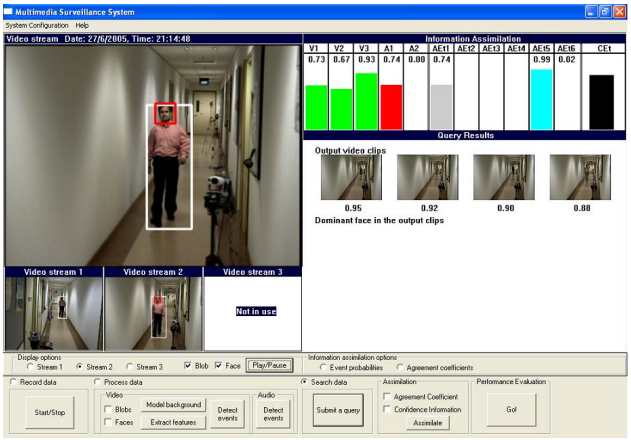


Figure 3: Multimedia Surveillance System

4.1 Data set

For our experiments, we have used data of more than twelve hours which has been recorded using the system consisting of two video cameras (Canon VC-C50i) and two USB microphones in the corridor of our school building. Over the period of more than twelve hours, a total of 92 events occurred over for a period of 1268 seconds. The details of various events and their time durations are given in Table 2. The graduate students from our lab volunteered to perform these activities.

4.2 Performance evaluation

The evaluation of proposed framework is performed based on two tasks - event detection and event classification. The evaluation of event detection task is characterized by two metrics - False Rejection Rate (FRR) and False Acceptance Rate (FAR). FRR is the ratio of number of events not detected to the total number of events, and FAR is the ratio of number of non-events detected to the total number of non-events. An event here refers to the observation made over a t_w period of time (Refer to section 3.2).

The event classification task is evaluated based on the accu-

Table 2: The data set

Events	Time duration (In seconds)
Standing	139
Walking	798
Running	142
Standing, Talking	30
Standing, Shouting	11
Standing, Knocking	59
Walking, Talking	80
Walking, Shouting	9

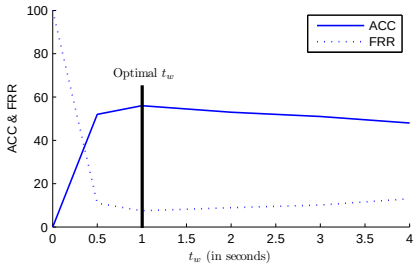


Figure 4: Determining the optimal value of t_w

racy (ACC) in classification. The metric ACC is defined as the ratio of number of events corrected classified to the total number of events that are detected to be the valid events. Again, an event refers to the observation made over a t_w time period.

As described in section 3.2, it is critical to determine the value of t_w. We have determined through experiments the suitable value of t_w to be 1 second. As can be seen from figure 4, at $t_w = 1$ second, we obtain the maximum accuracy (ACC) and minimum FRR.

4.3 Preprocessing steps

4.3.1 *Event detection in video streams*

The video is processed to detect human motion (running, walking and standing). Video processing involves two major steps - background modeling and blob detection. The background is modeled using an adaptive Gaussian method [Stauffer and Grimson 1999]. The blob detection is performed by first segmenting the foreground from the background using simple 'matching' on the three RGB color channels, and then using the morphological operations (erode and dilation) to obtain connected components (i.e. blobs). The matching is defined as a pixel value being within 2.5 standard deviations of the distribution. A summary of the video features used for various classification tasks is provided in Table 3(a). We assume that the blob of an area greater than a threshold corresponds to a human. The detected blob and its bounding rectangle is shown in figure 5. Once we compute the bounding rectangle (x, y, w, h) for each blob, where (x, y) denotes the top-left coordinate, w is the width and h is the height; we map the point $(x+w/2, h)$ (i.e. approximating with human's feets) in the image to a point (Ex, Ey) in 3-D world (i.e. on the corridor's floor), as shown in figure 6. To achieve this mapping, we calibrate the cameras and obtain a transformation matrix that maps image points to the points on corridor's floor. This provides the exact ground location of human in the corridor at a particular time instant.

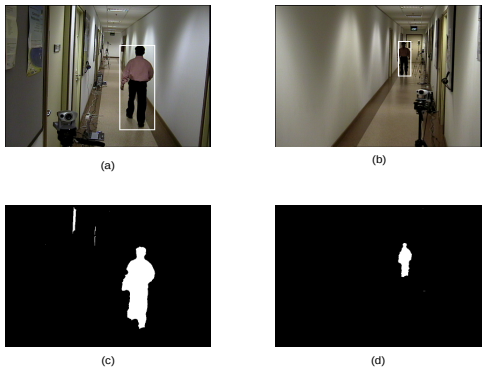


Figure 5: Blob detection in Camera 1 and Camera 2: (a)-(b) Bounding rectangle, (c)-(d) Detected blobs

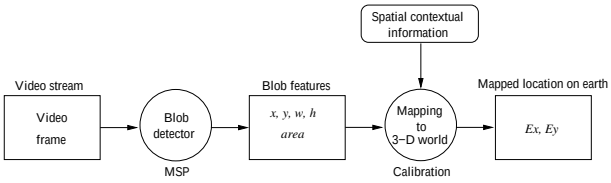


Figure 6: The process of finding from a video frame the location of a person on the corridor ground in 3-D world

The system identifies the start and end of an event in video streams as follows. If a person moves towards the camera, the start of event is marked when the blob's area becomes greater than a threshold and the event ends when the blob intersects the image plane. However, if the person walks away from the camera, the start and end of the event is inverted. The event detection is performed at regular time intervals of $t_w = 1$ second. Using the actual location of the person on the corridor's ground at the end of each time interval t_w, we compute the *average distance* travelled by a person on the ground. Based on this average distance, a Bayes classifier is first trained and then used to classify an atomic-event to be one of the classes - standing, walking and running.

4.3.2 *Event detection in audio streams*

Using the audio streams, the system detects events such footsteps, talking, shouting and door knocking. The audio (of 44.1 MHz frequency) is divided into the "audio frames" of 50 ms each. The frame size is chosen by experimentally observing that 50 ms is the minimum period during which an event such a footstep can be represented. We adopted a hierarchical (top-down) approach to model these events using a mixture of Gaussian (GMM). The top-down event modelling approach works better compared to the single-level multi-class modelling approach. We performed a separate study to find the suitability of features for de-

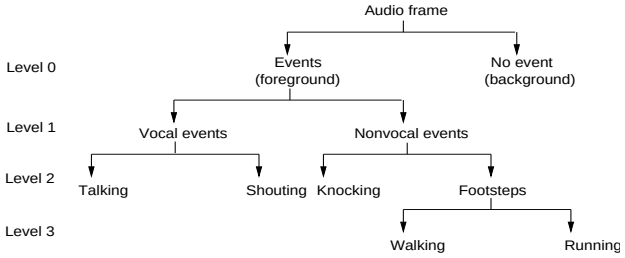


Figure 7: Audio event classification

Table 3: A summary of the features used for various classification tasks in video and audio streams

(a) Video

Classification task	Feature used
Foreground/Background	RGB channels
Running/Walking/Standing	Blob's displacement

(b) Audio

Classification task	Feature used
Foreground/Background	LFCC
Vocal/Nonvocal	LFCC
Talk/Shout	LPC
Footsteps/Door knocking	LFCC

tecting these audio events [Atrey et al. 2006]. Table 3(b) summarizes the audio features used for foreground/ background segmentation and for classification of events at different levels. The feature Log Frequency Cepstral Coefficients (LFCCs) with 10 coefficients and 20 filters worked well for foreground/background segmentation and for distinguishing between vocal/nonvocal and footsteps/knocking events. The LFCCs are computed by using logarithmic filter bank in frequency domain [Maddage 2006]. The Linear Predictor Coefficient (LPC) that have been widely used in speech processing community worked well for demarcating between talking and shouting events.

The Gaussian Mixture Model (GMM) classifier is employed to classify every audio frame (of 50 ms) into the audio events at different levels as shown in figure 7. At the top level (0), each input audio frame is classified as the foreground or the background. The background is the environment noise which represents 'no event' and is ignored. The foreground that represents the events, are further categorized into two classes - vocal and nonvocal (level 1). At the next level (2), both vocal and nonvocal events are further classified into "talking/shouting" and the "footsteps/door knocking" events, respectively. Finally, at the last level (3), the footsteps sequences are classified as "walking" or "running" based on the frequency of their occurrence in a specified time interval.

Similar to the video, the system makes a probabilistic decision about the events based on audio streams after every $t_w = 1$ second. Note that, in 1 second, we obtain 20 audio frames of 50 ms each. The audio event classification for the audio data of t_w time period is performed as follows. First, the system learns via training the number of audio frames corresponding to an event in the audio data of t_w time period. Then, a Bayesian classifier is employed to estimate the probability of occurrence of an audio event at a regular time

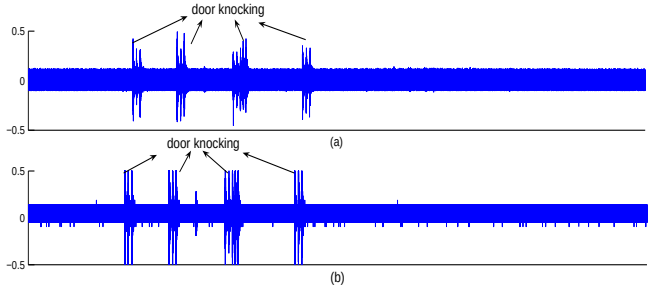


Figure 8: Audio data captured by (a) microphone 1 and (b) microphone 2 corresponding to the event $\mathbf{E}_k$

interval t_w.

4.4 Example of an event

In this section, we describe with an example how the proposed framework works in order to detect an event over a timeline. Let us consider a compound event $\mathbf{E}_k$ "A person is walking, knocking the door and then continued walking in the corridor". This event consists of atomic events occurring in two different ways. First, it consists of two atomic events occurring together i.e. "standing" and "door knocking" events. Second, it also consists of atomic events occurring one after another i.e. "walking" event followed by "standing/door knocking" event and then followed by "walking" event. The audio data captured using using microphone 1 and microphone 2 is shown in shown in figure 8. Some of the video frames captured by camera 1 and camera 2 corresponding to the event $\mathbf{E}_k$ and the bounding rectangles of the detected blobs in them are shown in figure 9.

The system detects the walking event using both audio and video streams, while standing and knocking events are detected based on video and audio streams, respectively. The probabilistic decisions about these atomic events are obtained based on respective streams at every $t_w = 1$ second. The overall decision for compound events are obtained along the timeline by assimilating the probabilistic decisions for atomic events as shown in figure 10. Note that in figure 10, the legends denote as follows: '$\circ$' - "standing", '$\square$' - "walking", '$\bigtriangledown$' - "running" and '$*$' - "door knocking" events.

Figures 10a-10d show the timeline-based probabilistic decisions based on individual streams. Figures 10e-10h show the combined decision about the event at a regular time interval with and without using streams' agreement/disagreement and confidence information.

It is interesting to note from figure 10 that using agreement coefficient though improves the accuracy of computing the probability of occurrence of an event, it is also important to use the confidence information to avoid incorrect results. For instance, using the stream's confidence information helps in obtaining correct results at time instants 3 and 4 in figure 10g-10h compared to the results at the same time instants in figure 10e-10f where confidence information is not used and an "walking" event is detected as "running". Note that the correct sequence of event is as follows: Time

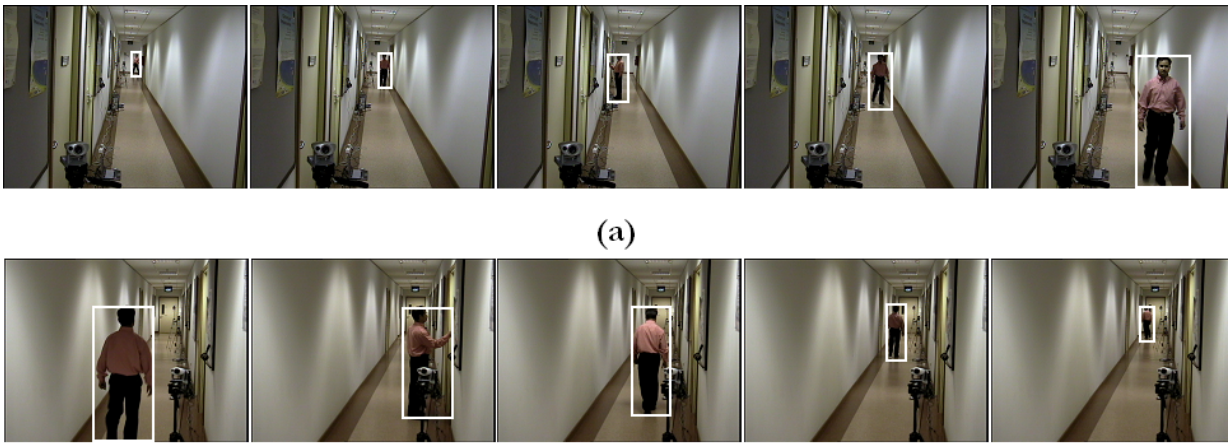


Figure 9: Some of the video frames captured by (a) camera 1 and (b) camera 2 corresponding to the event E

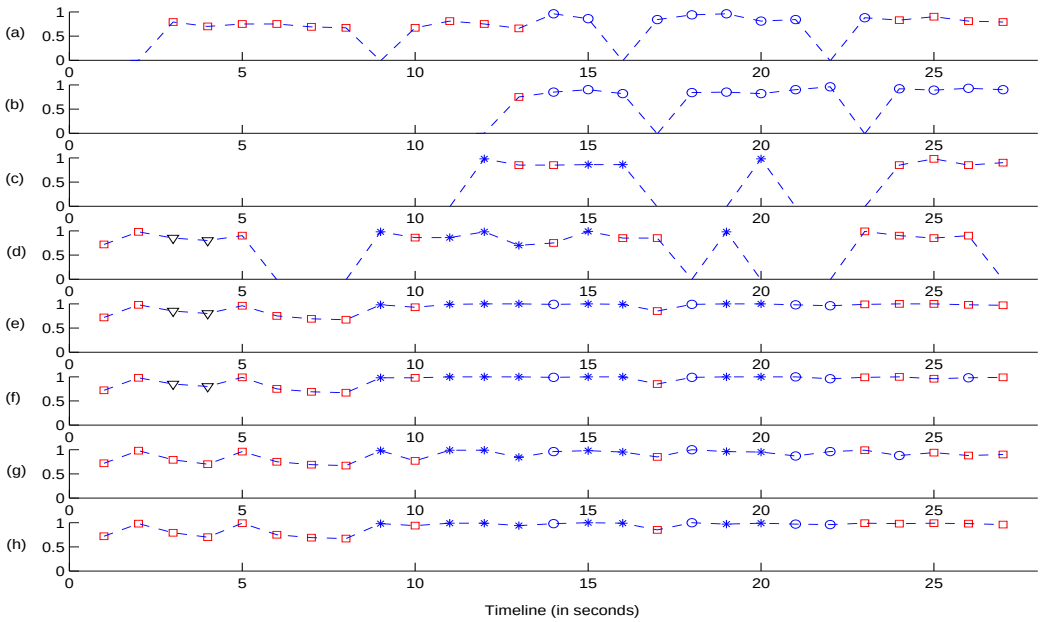


Figure 10: Timeline-based assimilation of probabilistic decisions about the event E. The legends denote the probabilistic decisions based on (a) Video stream 1 (b) Video stream 2 (c) Audio stream 1 (d) Audio stream 2 (e) All the streams (without agreement coefficient and confidence information) (f) All the streams (with agreement coefficient but without confidence information) (g) All the streams (with confidence information but without agreement coefficient) (h) All the streams (with both agreement coefficient and the confidence information)

Table 4: Results: Using individual streams with $Th = 0.70$

Stream	FRR	FAR	ACC
Video stream 1	0.12	0.01	0.60
Video stream 2	0.10	0.03	0.60
Audio stream 1	0.07	0.19	0.55
Audio stream 2	0.06	0.27	0.51

instants 1-9 "walking", 10-20 "standing/door knocking" and 21-27 "walking".

4.5 Overall performance analysis

4.5.1 Using Individual Streams

First, we performed event detection and classification using individual streams. The probability threshold Th value for determining the occurrence of an event was set to 0.70. The probability threshold Th is a threshold to convert a probabilistic decision into a binary decision (Refer to section 3.3.3). By comparing with the ground truth, we found the results as shown in Table 4. As can been from Table 4, FRR in video streams is higher than that in audio streams. This is because the video cameras were placed in such a way that they could not cover the whole corridor, and hence could not detect events outside their coverage area. On the other hand, since the microphones could capture the ambient sound even beyond the corridor area, they were able to detect the events those did not occur in the corridor region. Therefore, the microphones are found to have the FAR higher than that of video streams.

Using our whole set of events, we computed the accuracies (ACC) of event classification for all the four streams. We found the accuracy of individual streams to be moderate. However, it was found that the accuracy of event classification based on video streams was slightly better than that based on audio streams. We used these accuracy values to assign the confidences in all the four streams. Note that the overall accuracies of video streams is based on three types of events - "standing", "walking" and "running", while the audio streams' overall accuracies are determined based on five types of events - "walking", "running", "talking", "shouting" and "door knocking".

4.5.2 Assimilation of all streams

We performed assimilation of the probabilistic decisions obtained from individual streams in four different ways based on whether or not to use the agreement/disagreement information and the confidence information about them. The results are shown in Table 5. Note that these results are obtained by setting probability threshold Th and minimum time period t_w to 0.70 and 1 second, respectively.

Overall observations from Table 5 are as follows -

- Using multiple streams together provides better overall accuracy (ACC) and the reduced False Rejection Rate (FRR) as can be seen in the option 1 in Table 5. FAR is not evaluated in case of assimilating all the streams; since in the assimilation process, only the evidences of occurrence of the events are used, and therefore it does not affect FAR.

Table 5: Results: Using all the streams with $Th = 0.70$

Option	Agreement coefficient	Confidence information	FRR	ACC
1	No	No	0.011	0.72
2	Yes	No	0.011	0.78
3	No	Yes	0.010	0.76
4	Yes	Yes	0.012	0.80

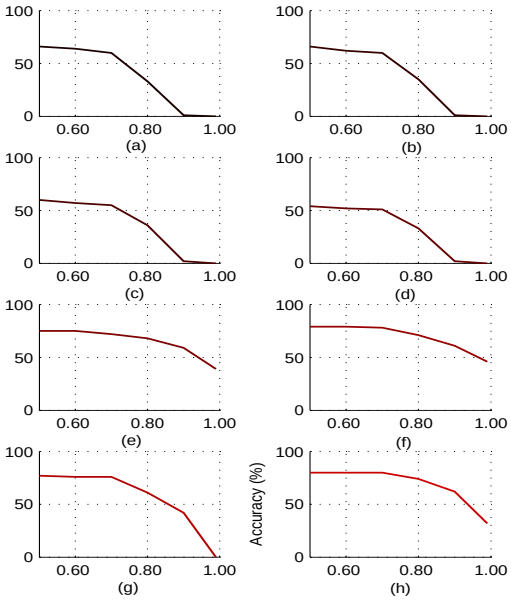
- The results (Table 5) imply that using agreement/ disagreement information among the streams is advantageous in obtaining more accurate results, however, using confidence information with it can further improve the overall accuracy of event detection and classification. Again, note that, the overall accuracies reported in Table 5 are for all the events listed in Table 2.

4.5.3 Early vs late thresholding

We also observed the accuracy of event classification by varying the probability threshold Th from 0.50 to 0.99. The results are shown in figure 11. Figure 11 shows how accuracy (ACC) decreases as the probability threshold Th increases for individual streams and for all streams when assimilated with four different options based on whether or not agreement coefficient and confidence information is used. It can be clearly seen from figure 11 that assimilation of all streams provide better accuracy even with a higher threshold, while individual streams fail in this respect. The accuracy decreases slowly for the combined evidences compared to the individual evidences. This implies that using agreement/disagreement among and confidence information of the streams in the assimilation process not only improves the overall accuracy, it also improves the accuracy of computing the probability of occurrence of the events. It also shows that early thresholding of the probabilistic decisions obtained based on individual streams leads to lesser accuracy; for example, in figure 11, at the probability threshold is 0.80, we obtain higher accuracies - 68, 71, 61 and 74 in the figures 11e-11h, respectively, after the assimilation of all streams compared to the accuracies - 33, 35, 36 and 33 in the figures 11a-11d, respectively, obtained using individual streams.

5. CONCLUSIONS

In this paper, we have presented a novel framework for assimilation of information in order to detect events in the surveillance and monitoring systems that utilize multifarious sensors. The experimental results have shown that the use of agreement coefficient among and the confidence information of media streams helps in obtaining more accurate and credible decisions about the events. The results have also shown that the False Rejection Rate for event detection can be significantly reduced using all the streams together. In future work, there are many other issues which need to explored such as - first, how to determine the minimum time period to confirm different events; second, it would be interesting to see how framework will work when the information from different sources would be made available at different time



x-axis: Probability Threshold (Th), y-axis: Accuracy (ACC)

Figure 11: Plots: Probability Threshold vs Accuracy. (a) Video stream 1 (b) Video stream 2 (c) Audio stream 1 (d) Audio stream 2 (e)-(h) All streams after assimilation with the four options given in Table 5

instances, what would be the ideal sampling rate of event detection and information assimilation; and finally, how the confidence information about a stream (newly added in the system) can be computed over time using its agreement/ disagreement with the other streams whose confidence information are known, and how it would evolve over time with the changes in environment.

6. REFERENCES

Atrey, P. K., Kankanhalli, M. S., and Jain, R. 2005. Timeline-based information assimilation in multimedia surveillance and monitoring systems. In *The ACM International Workshop on Video Surveillance and Sensor Networks*. Singapore, 103–112.

Atrey, P. K., Kankanhalli, M. S., and Oommen, J. B. 2006. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Transactions on Multimedia Computing, Communications and Applications*. (To appear).

Atrey, P. K., Maddage, N. C., and Kankanhalli, M. S. 2006. Audio based event detection for multimedia surveillance. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Toulouse, France, V813–816.

Benediktsson, J. A. and Kanellopoulos, I. 1999. Classification of multisource and hyperspectral data based on decision fusion. *IEEE Trans. on GeoScience and Remote Sensing 37,* 3 (May), 1367–1377.

Bloch, D. A. and Kraemer, H. C. 1989. 2 × 2 Kappa coefficients: Measures of agreement or association. *Journal of Biometrics 45,* 1, 269–287.

Chair, Z. and Varshney, P. R. 1986. Optimal data fusion in multiple sensor detection systems. *IEEE Transactions on Aerospace and Electronic Systems 22*, 98–101.

Checka, N., Wilson, K. W., Siracusa, M. R., and Darrell, T. 2004. Multiple person and speaker activity tracking with a particle filter. In *International Conference on Acoustics Speech and Signal Processing*.

Chieu, H. L. and Lee, Y. K. 2004. Query based event extraction along a timeline. In *International ACM SIGIR Conference on Research and development in Information Retrieval*. Sheffield, UK, 425–432.

Genest, C. and Zidek, J. V. 1986. Combining probability distributions: A critique and annotated bibliography. *Journal of Statistical Science 1,* 1, 114–118.

Hershey, J., Attias, H., Jojic, N., and Krisjianson, T. 2004. Audio visual graphical models for speech processing. In *IEEE International Conference on Speech, Acoustics, and Signal Processing*. Montreal, Canada, V:649–652.

Kam, M., Zhu, Q., and Gray, W. S. 1992. Optimal data fusion of correlated local decisions in multiple sensor detection systems. *IEEE Transactions on Aerospace and Electronic Systems 28,* 3 (July), 916–920.

Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Journal of Biometrics 45,* 1, 255–268.

Maddage, N. C. 2006. Content based music structure analysis. Ph.D. thesis, School of Computing, National University of Singapore.

Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphye, K. 2002. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing 11*, 1–15.

Nock, H. J., Iyengar, G., and Neti, C. 2002. Assessing face and speech consistency for monologue detection in video. In *ACM International Conference on Multimedia*.

Rao, B. S. and Whyte, H. D. 1993. A decentralized bayesian algorithm for identification of tracked objects. *IEEE Transactions on Systems, Man and Cybernetics 23*, 1683–1698.

Siegel, M. and Wu, H. 2004. Confidence fusion. In *IEEE International Workshop on Robot Sensing*. 96–99.

Stauffer, C. and Grimson, W. E. L. 1999. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2. Ft. Collins, CO, USA, 252–258.

Wu, Y., Chang, E. Y., Chang, K. C.-C., and Smith, J. R. 2004. Optimal multimodal fusion for multimedia data analysis. In *ACM International Conference on Multimedia*. New York, USA, 572–579.