ABSTRACT

Most surveillance and monitoring systems nowadays utilize multiple types of sensors. 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 hierarchical probabilistic method for information assimilation in order to detect events of interest in a surveillance and monitoring environment. The proposed method adopts a bottom-up approach and performs assimilation of information at three different levels - media-stream level, atomic-event level and compound-event level. To detect an event, our method uses not only the current media streams but it also utilizes their two important properties - first, accumulated past history of whether they have been providing the concurring or contradictory evidences, and - second, the system designer’s confidence in them. A compound event, which comprises of two or more atomic-events, is detected by first estimating probabilistic decisions for the atomic-events based on individual streams, and then by aligning these decisions along a timeline and hierarchically assimilating them. The experimental results show the utility of our method.

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

As a result of increasing public security threats, majority of developed cities around the world are now being equipped with thousands of sensors including video camera and even microphones [1] with a primary goal of monitoring and recording events of interest as and when they occur in the environment under surveillance and monitoring. In the current-generation surveillance systems [18], where a number of asynchronous and multifarious sensors are employed, assimilation of the information obtained from them in order to infer the events from the environment is an important and challenging research problem. Information assimilation refers to the process of combining the sensory and non-sensory information using the context and past experience. The issue of information assimilation is important because the information obtained from multiple sources when assimilated provides more accurate inferences about the environment than the individual sources. The information assimilation is challenging because of the following reasons -

1. Asynchrony and diversity of sensors: The sensors provide data\(^1\) of different formats and at different rates. For example, a video sensor provides frames at a rate which could be different from the rate at which audio samples are obtained, or even two video sensors can provide frames at different rates. The non-sensory information (e.g. past-record of a criminal person) can also be in a different format. Assimilation of information from the asynchronous and diverse sources requires its alignment along a “timeline”, which is challenging. Timeline refers to a measurable span of time with information denoted at key points.

2. Agreement/disagreement among media streams: The sensors capturing the same environment provide concurring or contradictory evidences about what is happening in the environment. This agreement/disagreement information among the media streams can be used to strengthen the overall decision about the events happening in the environment. For example, if two sensors have been providing the concurring evidences in the past, it makes sense to give more weight to their current combined evidence compared to the case

\(^1\)In context of multimedia, we call sensor data as “media stream”
if they provided contradictory evidences in the past [16]. Therefore, how to utilize the past history of agreement/disagreement information among the media streams is a challenging issue.

3. **Confidence in media streams**: The designer of a surveillance and monitoring system can have different confidence levels in different media streams while detecting different events. The confidence in a media stream is directly related to its accuracy. For example, if an event is 70% times correctly detected based on a media stream, one can have 70% confidence in it. Moreover, it makes more sense to give a higher weight to the media stream which has a higher confidence level. Note that the accuracy of a media stream includes the measurement accuracy of sensor as well as the accuracy of the algorithms used to process the media stream. The accuracy can be learned through experiments. Other issue which arises is the fusion of confidence levels. We illustrate this with an example. Let we have two media streams whose confidence levels are 70% and 80%. Now, at the current instant, if the two streams are agreeing about an event, we must believe more in the decision based on combined streams than what we would believed in them individually. On the other hand, if the streams disagree, the confidence in the combined decision would decrease [16]. This is clearly an issue of how to fuse individual confidences in two different streams to obtain the overall confidence in a group of these two streams. These issues are of extreme importance and significant challenge.

In this paper, we propose a hierarchical probabilistic method for information assimilation in order to detect events in surveillance and monitoring systems that utilize multiple disparate sources. The proposed method assimilates the media streams using their agreement/disagreement/confidence information. 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. The confidence in each stream is computed based on how accurate it has been in the past. We also propose the methods for fusing the agreement coefficients and for fusing the confidence levels of media streams. Our method performs multimedia information assimilation at three different hierarchical levels - media-stream level, atomic-event level and compound-event level. At media-stream level, we do early assimilation of features that are extracted from the media stream; and at atomic-event and compound-event levels, we employ late integration strategy by assimilating the decisions about atomic-events and compound-events, respectively.

To distinguish among the events, compound-events and the atomic-events, we define them 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 at a location over a period of time.

![Figure 1: Timeline alignment of the decisions obtained based on various media streams](image)

**Definition 3.** Compound-event is union of two or more atomic-events.

An event could be a compound-event or simply an atomic-event. If an event is a compound-event, we first decompose it into its constituent atomic-events. For instance, a compound-event “a person is running and shouting in the corridor” is decomposed into following two atomic-events - “a person is running in the corridor” and “a person is shouting in the corridor”. We do this decomposition because different atomic-events can be detected using different types of media streams, and therefore it makes sense to identify the data-level interpretation of the domain-level atomic-events. The domain and data levels refer to the high-level semantics and low-level features, respectively. The domain level atomic-event “a person is running in the corridor” is interpreted at data level as “a blob of certain size is displaced in the image plane at certain rate from one frame to another frame”. Similarly, the atomic-event “a person is shouting in the corridor” is interpreted as “rate of change of energy in a sound signal is greater than a threshold”.

The proposed method adopts a bottom-up approach which goes as follows. It first estimates along a timeline the probability of the occurrence of an atomic-event within the specified time-interval by employing the event detectors on individual media streams. The event detector assimilates the features of that stream - early assimilation at media-stream level. Early assimilation at the media-stream level makes sense because it is good to use all the possible features of a media stream in order to decide about the atomic-event. Then at atomic-event level, we align along a timeline the individual decisions about the atomic-event obtained based on relevant media streams as shown in figure 1, and estimate the overall probability of the occurrence of atomic-event by assimilating the individual decisions (in the form of probabilities) using a Bayesian approach [2] - late assimilation at atomic-event level. Once the probabilistic decisions about all the atomic-events are obtained, assimilation of these decisions is finally performed at compound-event level to obtain the overall probability of the occurrence of compound event - late assimilation at compound-event level. The late assimilation at atomic-event level and compound-event level makes sense because the media streams, using which the independent decisions about (atomic) events are obtained, are of different types. Hence it is more reasonable to assimilate their decisions rather than the heterogenous media features.

Timeline synchronization of these decisions about an atomic-event is important as well as challenging because of the fact that the system provides decisions based on different media...
streams at different times, and the assimilation of decisions is performed only at the key points in the timeline. Identifying these key points is clearly an issue.

We argue the advantage of our approach of assimilating probabilistic decisions over the approach of integrating the binary decisions as follows. The approaches, that integrate binary decisions, always contain some errors due to thresholding. For example, let an event detector finds 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 imply 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. On contrary to the early thresholding, we argue for late thresholding. We do not obtain binary decisions at this level, but assimilate using our proposed method these probabilistic decisions and obtain the overall probability (which is usually more than the individual probabilities e.g. 0.85 in this case). The thresholding of this probabilistic decision at late stage is less erroneous.

Multimedia information assimilation is different from multimedia 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 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 propose a hierarchical probabilistic method for information assimilation in order to detect events in a multimedia surveillance and monitoring environment that utilizes multiple heterogeneous sensors. Moreover, we have introduced the notion of agreement/disagreement among the media streams and established its importance in the assimilation process. We have formulated the computation and fusion of the agreement coefficients among the streams. The proposed method for information assimilation also integrates the confidence information of media streams. We have proposed a method for confidence fusion. We have shown how the probabilistic method for information assimilation provide along a timeline more accurate decisions about the events in the surveillance and monitoring environment.

Rest of this paper is organized as follows. In section 2, we discuss the related work. We present our method 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

Multimedia researchers have used early fusion strategy by combining video and audio at feature-level for various problems including speech processing [9] and recognition [13], tracking [7], and monologue detection [14] 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.

An increasing amount of work is also reported in sensor fusion literature where a global decision is made by fusing the local decisions obtained from each sensor. [15] presented a sensor fusion algorithm for identification of tracked targets in a decentralized environment. [6] established an optimal fusion rule with the assumption that each local sensor made a predetermined decision and each observation was independent. [11] generalizes their solution for fusing the correlated local decisions.

Similar to [19], 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. upgrading or graceful degradation) in terms of media streams used in the assimilation process [3]. Note that, in late assimilation, we consider the media streams to be “decision-wise correlated”.

The works cited above do not explicitly compute and utilize the correlation information among the sensors. In contrast, our method computes the correlation information (we call it to be “agreement coefficient”) based on how concurring or contradictory are the evidences from media streams. Intuitively, higher the agreement among the streams, more would be the confidence in global decision, and vice versa [16]. In the past, various forms of correlation coefficients have been studied and used for diverse applications. But, most of them 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 [12] and Kappa coefficient [5] can’t be used in our case since they are evaluated to zero when the covariance among the observations is zero. Therefore, we model 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 work in decision fusion literature does not consider the notion of having confidences in the decisions. In our method, we incorporate the stream’s confidence information. [16] has also recently 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 model the confidence fusion by using a Bayesian formulation because it is both simple and computationally efficient [15].

3. PROPOSED METHOD

In order to detect an event in a surveillance and monitoring environment, the proposed method adopts a bottom-up approach and performs information assimilation at three hierarchical levels - media-stream level, atomic-event level and the compound-event level. The work flow of the method is depicted in figure 2. Let a surveillance and monitoring system consists of a heterogeneous media sensors capturing the data from environment. We employ $n$ Media Stream Pro-
cessors (MSP to MSPn), where each MSPi, 1 ≤ i ≤ n, is a set of media processing tools that extracts features from the media stream Mi; for example, a blob detector extracts blob from a video stream. The features extracted from each media stream are stored in their respective databases. A compound-event E, which comprises of two or more atomic-events ei, (1 < j ≤ r, r being the maximum possible of atomic-events in a compound event), is detected hierarchically. First, atomic-events are detected using the relevant media streams, and then these decisions are assimilated hierarchically to obtain an overall decision for the compound event E, as described in the subsequent subsections. To further illustrate, we provide the following example.

Example 1: Let us consider the compound-event E is “A person is walking and shouting in the corridor”. This compound-event is composed of two atomic-events: e1 = “A person is walking in the corridor” and e2 = “A person is shouting in the corridor”. The atomic-event e1 can be detected using two types of media sensors available in the environment (i.e. geometry of the monitored space, location, orientation and coverage space etc of media sensors) is also used whenever required by the event detectors. The event detectors provide their decisions in probabilities in the timeline, we iteratively integrate all the media streams using a Bayesian approach. The proposed approach allows for incremental and iterative addition of new information. Let P(ei|Mj) denote probability of occurrence of atomic-event ei at time t based on from media streams Mi, Mj, …, Mj−1. The updated probability P(ei|Mj) (i.e. the overall probability after assimilating the new media stream Mi) can be recursively computed as -

\[ P(e_i|M_i^t) = \frac{P(M_i|M_{j-1})P(e_i|M_j^t)}{P(M_i|M_{j-1})} \]

where, \( \alpha_s \) is a normalizing constant.

The 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. 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 [4]. 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 we have 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 express quantitatively the goodness of each source. The rule is formulated as: T(p1, p2, …, pn) = \( \sum_{i=1}^{n} w_i p_i \) where, \( p_i, 1 ≤ i ≤ n \), are the individual probabilistic decisions; and \( w_i, 1 ≤ i ≤ n \) 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 assim-
aspects of an event - the occurrence and non-occurrence of decisions based on two sources

posed assimilation model. The formulation is shown as -

\[
P_i = \frac{1}{2} \left[ \left( P_{i-1} \right)^{f_i} \left( 1 - P_{i-1} \right)^{1-f_i} \right] \sigma_i^n \cdot \left( 1 - P_{i-1} \right) \left( 1 - P_{i-1} \right) \left( P_{i-1} \right)^{f_i} \left( 1 - P_{i-1} \right)^{1-f_i} \cdot \sigma_i^{n-1},
\]

where, \( P_i = P(e_{ij} | M_i) \) and \( P_{i-1} = P(e_{ij} | M_i^{i-1}) \) are the probabilities of atomic-event \( e_j \) being occurring using \( M^i \) and \( M^{i-1} \), respectively, at time instant \( t \). \( p_i = P(e_{ij} | M_i, t) \) is probability of the occurrence of atomic-event \( e_j \) based on only \( i^{th} \) stream at time instant \( t \). Similarly, \( F_{i-1} + f_i = 1 \) is the confidence in \( M^{i-1} \) and \( M_i \), respectively. The computation of confidence for a group of media streams is described in section 3.2.3. The \( \tau_i \in [-1, 1] \) is the agreement coefficient between two sources \( M^{i-1} \) and \( M_i \). The limits -1 and 1 represent full disagreement and full agreement, respectively, between the two sources. The modelling of \( \tau_i \) is described in section 3.2.2.

### 3.2.2 Modelling of the agreement coefficient

Let the measure of agreement among the media streams at time \( t \) be represented by a set \( \Gamma(t) \) which is expressed as:

\[
\Gamma(t) = \{ \gamma_{ik}(t) \}
\]

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 \( t \).

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

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

where, \( p_i(t) = P(e_{ij} | M_i) \) and \( p_k(t) = P(e_{ij} | M_k) \) are the individual probabilities of occurrence of atomic-event \( e_j \) based on media streams \( M_i \) and \( M_k \), respectively, at time \( t \geq 1 \); and \( \gamma_{ij}(0) = 1 - 2 \times \text{abs}(p_i(0) - p_k(0)) \). 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 (\( \gamma_{ik} = -1 \)).

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

\[
\tau_i = \frac{1}{i-1} \sum_{s=1}^{i-1} \gamma_{si}
\]

where, \( \gamma_{si} \) for \( 1 \leq s \leq i - 1 \) and \( 1 < i \leq n \) 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 \( M^{i-1} \) 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 $\tau_i$ is used for combining $M_i$ with $M^{i-1}$ as described before in equation (5).

3.2.3 Confidence fusion

In the context of streams, we relate confidence in a media stream to its accuracy. We compute the accuracy of a media stream by determining how many times an event is correctly detected based on it out of the total number of tries.

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. For example, consider the case when a video stream has 70% confidence level and an audio stream has 60% confidence level. What would our confidence be in a group which contains both video and audio stream? 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 $M_1$ and $M_2$ is computed as follows:

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

where $f_i$ and $f_k$ are the individual confidence levels of media streams $M_1$ and $M_2$, respectively. In the above formulation, although the media streams are correlated in their decisions, we assume that they are mutually independent in terms of their confidence levels. The $f_{ik}$ is clearly the joint probability that the system designer is interested in. 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 Compound-event level assimilation

At the compound-event level, we compute the overall probability $P_E$ of the occurrence of compound-event $E$ by assimilating the probabilistic decisions $p_{ij}$, $1 < j \leq r$ about the $r$ atomic-events by using the following assimilation model -

$$P_E = \frac{\prod_{j=1}^{r} p_{ij}}{\prod_{j=1}^{r} p_{ij} + \prod_{j=1}^{r} (1-p_{ij})} \quad (11)$$

If $P_E$ is found greater than the threshold $T_h$, the system decides in favor of the occurrence of compound event $E$, else it decides against it.

Since the atomic-events are independent, the agreement coefficients among them are considered as zero, and hence it is not integrated into equation (11). For example, atomic-events $e_1$ = “A person is walking in the corridor” and $e_2$ = “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 related to media streams not to the atomic-events.

4. EXPERIMENTAL RESULTS

To demonstrate the utility of our information assimilation method, we present experimental results in a surveillance and monitoring scenario of the corridor in our school building. We installed two video sensors (cameras) and one audio sensor (microphone) in the corridor as shown in the layout given in figure 3. In figure 3, we show their location and orientation. The objective is to monitor human activities such as human motion (viz “walking”, “running”, and “standing”), and the unusual sounds such as “shouting” or “noisy footsteps” etc. The video sensors $M_1$ and $M_2$ cover the whole corridor from opposite sides. The audio sensor $M_3$ is placed in the corridor to capture the ambient sound. To detect the events of human’s running, walking and standing, we use data from two video cameras and one microphone; while unusual sounds are detected based on only audio sensor data.

We consider the compound event $E$ - “A person is walking and shouting in the corridor” which comprises of two atomic-events $e_1$ and $e_2$ as illustrated in Example 1.

4.1 Atomic-event detection in video streams

To detect the human motion in the corridor, we have used an adaptive Gaussian method [17, 10] to model each pixel as a mixture of Gaussian. Once the background is modelled, we segment the foreground (blobs) from it using simple ‘matching’ on the three RGB channels in video frames. The matching is defined as a pixel value being within 2.5 standard deviation of the distribution. In the extracted foreground, we perform morphological operations (erode and dilation) to obtain connected components (i.e. blobs). We assume that the blob of an area greater than a threshold corresponds to a human. This assumption is reasonable since the focus of this paper is not the object or human recognition, but it is the information assimilation. As shown in figure 4, we extracted from each video frame 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 of bounding rectangle of the blob. We then map the point \((x + w/2, h)\) (i.e. approximating with human’s foots) in the image to a point \((Ex, Ey)\) in 3-D world (i.e. on the corridor’s ground). To achieve this mapping, we calibrate the cameras and obtain a transformation matrix that maps image points to the points on corridor’s ground. This provides the exact ground location of human in the corridor at a particular time instant.

Based on the presence of potential blobs in a sequence of video frames, we identify the atomic-events. For example, if a person moves towards the camera, the start of atomic-event is marked when the blob’s area becomes greater than a threshold and the atomic-event ends when the blob intersects the image plane. However, if the person walks away from the camera, the start and end of the atomic-event is inverted. Once an atomic-event is detected, we divide the time duration for which the event occurred into the time-inverted. Once an atomic-event is detected, we divide the time period to be assimilated. For example, the decision based on two video streams \((\text{E}_1, \text{E}_2)\) along with its probabilistic decisions \(p(e_1|M_1)\) (figure 5b) and \(p(e_2|M_2)\) (figure 6c) for the atomic-events \(e_1\) and \(e_2\), respectively is shown in figure 6. The probabilistic decisions obtained from the two video sensors and one audio sensor are aligned along a timeline. In order to derive an overall decision about an event, it is necessary to key out the periodic time instances when they provide their decisions. In our experiments, we have observed that \(t_w = 2\) seconds is the minimum time period in which the atomic-event \(e_1\) can be detected using a video stream, therefore we have chosen this time period to be \(t_w = 2\) seconds. Note that this minimum time period can be different for different atomic-events.

Figure 7 shows the assimilation of media streams without using their agreement/ disagreement and confidence information in order to detect compound event \(E\). First, two video streams (figure 7a) and then all three streams (figure 7b) are assimilated (at atomic-event level) to detect the atomic-event \(e_1\). Finally, as shown in figure 7c, the decisions for both the atomic-events \(e_1\) and \(e_2\) are assimilated (at compound-event level) in order to make an overall decision for the compound event \(E\). As can be seen in figure 7a (with two streams), from time 7.9275 ms onwards, probability of the occurrence of atomic-event \(e_1\) is more compared to that in figure 5b (camera 1) and in figure 5d (camera 2). Also, in figure 7b (with three streams), from time 7.9275 ms onwards, the probability of the occurrence of \(e_1\) is higher than that in figure 7a (with two streams). This shows that a higher number of concurring events strengthen the overall decision about an event. Note that the assimilation is performed only at those key points in timeline where the decisions based on more than one media stream are available.

We show the integration of streams’ agreement/ disagreement information in figure 8. Figure 8a shows how the agreement coefficient \((\gamma_{12})\) between the two video streams evolves over time. The agreement coefficient \(\gamma_{12}\) further strengthens the decision based on two video streams \(M_1\) and \(M_2\) as can be seen from the time 7.9275 ms onwards in figure 8b compared to that in figure 7b. In figure 8c, we show the fused agreement coefficient \((\gamma_{123})\) among three streams (two video streams and one audio stream) (refer to equation (8)) which also evolves over time. The \(\gamma_{123}\) implies the agreement coefficient between audio stream and the group of two video streams. Figure 8d shows along a timeline the overall probability of the occurrence of atomic-event \(e_1\) based on all three streams. Finally, the figure 8e shows the overall decisions along a timeline for the compound event \(E\). As can be seen in figure 8d (for atomic-event \(e_1\)) and in figure 8e (for compound-event \(E\)), the overall probabilistic decisions about the event is closer to 1 which shows that the use of agreement coefficient among streams minimizes the errors in the derived decision. Note that in this case, we assumed that all the streams bear equal confidence levels.

The effect of integrating the confidences in streams is shown in figure 9. We have first computed the confidence levels of the three media streams by running the experiments for 15 events of walking, standing, running and shouting. By comparing the results with the ground truth, we have found the the accuracy of the three streams as follows - video stream \(M_1\): 0.75, video stream \(M_2\): 0.70, and the audio stream \(M_3\): 0.60. The confidence information becomes...
Figure 5: Timeline-alignment of the probabilistic decisions obtained based on (b) video stream $M_1$ (shown in (a)) and (d) video stream $M_2$ (shown in (c)) for the atomic-event $e_1$.

Figure 6: Timeline-alignment of the probabilistic decisions obtained based on (a) audio stream $M_3$ for the atomic-events (b) $e_1$ and (c) $e_2$.

Figure 7: Timeline-based information assimilation for detecting a compound event $E$ - “A person is walking and shouting in the corridor” based on two video streams $M_1, M_2$ and one audio stream $M_3$ without using their agreement/disagreement and confidence information.
very important when the two streams provide contradictory evidences with equal probabilities. In such case, we choose the one with a higher confidence level. Figure 9a shows the overall confidence \( f_{12} \) in a group of two video streams \( M_1 \) and \( M_2 \). This is computed using equation (9). Similarly, the confidence in a group of two video streams is fused with the confidence level of one audio stream (using equation (10)) as shown in figure 9d. As expected, the confidence in a group of streams increases with the number of streams in it. The step-by-step hierarchical assimilation to make an overall decision for the compound event \( E \) is shown in figure 9b, in figure 9c and in figure 9e.

To sum up, the experimental results demonstrate that the use of media streams’ agreement/disagreement and confidence information provides more accurate decisions for the atomic-events and the compound-event. The late assimilation strategy performs well considering that fact it offers us flexibility to add or drop a stream which is highly essential in multi-sensor surveillance and monitoring systems.

5. CONCLUSIONS

In this paper, we have presented a novel method for assimilation of information in the surveillance and monitoring systems that utilize multifarious sensors. The experimental results have shown that the use of agreement coefficient among the confidence information of media streams helps in obtaining more accurate and credible decisions about the events. However, these results are preliminary and future work will be to perform more extensive experiments to evaluate the performance of our information assimilation method and compare it with the existing approaches of multimedia fusion. In future work, there are many other issues which need to explored such as - first, 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; second, how the confidence level of a media stream would evolve over time with the changes in environment; and third, the scalability analysis - how would adding or dropping a stream affect the accuracy of overall decision for an event. It would also be interesting to explore how high-level “concepts” (e.g. stampede) can be modelled using atomic-events and how it can be used through our method.

6. ACKNOWLEDGMENTS

The authors would like to thank Vinay Kumar and Anurag Kumar from IIT, Kharagpur for providing help in parts of the system implementation.

7. REFERENCES

Figure 8: Timeline-based information assimilation for detecting a compound event $E$ based two video streams $M_1, M_2$ and one audio stream $M_3$ using their agreement/disagreement but without using their confidence information.

Figure 9: Timeline-based information assimilation for detecting a compound event $E$ based two video streams $M_1, M_2$ and one audio stream $M_3$ using both their agreement/disagreement and the confidence information.