Confidence Building Among Correlated Streams in Multimedia Surveillance Systems

Pradeep K. Atrey[1], Mohan S. Kankanhalli[2], and Abdulmotaleb El Saddik[1]

[1] School of Information Technology and Engineering, University of Ottawa, Canada
{patrey, abed}@mcrlab.uottawa.ca
[2] Department of Computer Science, School of Computing, National University of Singapore, Republic of Singapore
mohan@comp.nus.edu.sg

Abstract. Multimedia surveillance systems utilize multiple correlated media streams, each of which has a different confidence level in accomplishing various surveillance tasks. For example, the system designer may have a higher confidence in the video stream compared to the audio stream for detecting humans running events. The confidence level of streams is usually precomputed based on their past accuracy. This traditional approach is cumbersome especially when we add a new stream in the system without the knowledge of its past history. This paper proposes a novel method which dynamically computes the confidence level of new streams based on their agreement/disagreement with the already trusted streams. The preliminary experimental results show the utility of our method.

1 Introduction

Current surveillance systems often utilize multiple types of sensors like microphones [1], motion detectors [2] and RFIDs [3] etc in addition to the video cameras. As different sensors have different capabilities of performing various surveillance tasks, the designer of a multimedia surveillance system usually has different confidence levels in the evidences obtained based on the data of dissimilar sensors (we call sensor's data to be the "media streams" from now onwards) for accomplishing various tasks. For instance, the system designer may have higher confidence in a video stream compared to an audio stream for detecting faces, and may also have high confidence in an audio stream for detecting talking/shouting events.

In order to accomplish any surveillance task, the system assimilates relevant media streams. As the different streams have different confidence levels associated for accomplishing different tasks, it is important to utilize the confidence information of streams in their assimilation by appropriately assigning the weights to them [4]. The confidence in a stream is related to its accuracy. The higher

T.-J. Cham et al. (Eds.): MMM 2007, LNCS 4352, Part II, pp. 155–164, 2007.

the accuracy of a stream, higher the confidence we would have in it. In the assimilation process, it makes sense to give more weight to a stream which has a higher confidence factor.

However, the computation of confidence information for each stream is cumbersome especially when we dynamically add the new streams to a multimedia surveillance system. The usual approach for determining the confidence in a stream is to first compute, in advance, its accuracy and then assign the confidence level to it based on its accuracy. This is often difficult because the system may provide different accuracies for different events when detected based on different media streams. Precomputation of accuracies of all the streams, that too for all events under different contexts, requires significant amount of training and testing, which is often tedious and time consuming. Moreover, for the streams which are added later in the system, there is no way to find their past accuracy. Therefore, it is important to devise a method to dynamically determine the confidence levels of streams without precomputing it.

In this paper, we propose a novel method for dynamically computing the confidences in a newly deployed stream based on the knowledge of the existing "trusted" stream(s) and the agreement coefficient among the newly deployed stream and the existing trusted streams. We call a stream to be "trusted" if its confidence level is greater than a threshold. The agreement coefficient between the streams is computed based on how agreeing or disagreeing the evidences obtained based on them have been in the past.

To illustrate our core idea, we provide the example of TV news channels. Let we follow a trusted CNN news channel. We also start watching an arbitrary XYZ news channel and compare the news content provided on both the channels. Over a period of time, our confidence in the XYZ channel will grow if the news content of both channels are found to be similar, and vice versa.

Rest of this paper is organized as follows. In section 2, we describe the related work. We formulate the problem of determining confidence in a stream in section 3. Section 4 presents our proposed method. We present the experimental results in section 5. Finally, section 6 concludes the paper with a discussion on the future work.

2 Related Work

In the past, the confidence has been used in the context of data management in sensor networks. Tatbul et al. [5] compute the confidence in a stream based on how it has helped in making the accurate decisions in the past. Tavakoli et al. [6] proposed a method for event detection that uses historical and spatial information in clusters in order to determine a confidence level that warrants a detection report with high confidence. Ioannou et al. [7] also employed a confidence-based fusion strategy to combine multiple feature cues for facial expression recognition. However, the works at [5], [6] and [7] did not elaborate on how the confidence value is used in the integration of information.

Siegel and Wu [8] has 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[4].

In all the past works, the confidence in streams has been computed based on their past accuracy. This work is different from the past works in that, our method computes the confidence level of streams based their agreement/ disagreement with the trusted streams. Agreement coefficient among streams is computed based on how concurring or contradictory evidences they provide. Agreement coefficient between any two streams is different from mutual information [9] between them in that the former connotes the measure of mutual consistency or contradiction between the two streams while the latter implies how much information does one stream convey about another one.

3 Problem Formulation

We formulate below the problem of determining the confidence level of a media stream:

$\mathcal{M}1$. $\mathsf{S}$is a multimedia surveillance system designed for detecting a set E of events, and it consists of $n \geq 1$ heterogeneous sensors that capture data from the environment. Let $\mathbf{M}^n = \{M_1, M_2, \ldots, M_n\}$ be the media streams obtained from n sensors.

$\mathcal{M}2$. For $1 \leq i \leq n$, let $0 < p_i(t) < 1$ be the *probability* of occurrence of an event based on individual i^{th} media stream at time instant t. The $p_i(t)$ is determined by first extracting the features from media stream i and then by employing an event detector (e.g. a trained classifier) on them. Also, let $P_\Phi(t)$ be the 'fused probability' of occurrence of the event at time t based on a subset $\Phi \in \mathcal{P}(\mathbf{M}^n)$ of media streams. The 'fused probability' is the overall probability of occurrence of the event based on a group of media streams [4].

$\mathcal{M}3$. For $1 \leq i \leq n$, let $0 < f_i(t) < 1$ be the system designer's *confidence* in the i^{th} stream at time instant t. The confidence in at least one media stream is learned by experimentally determining its accuracy. More the accurate results we obtain based on a stream, more the confidence we would have in it. Also, there exists a subset $T \subseteq \mathbf{M}^n$ (with $|T| \geq 1$) of streams in which the confidence level of streams is greater than or equal to a threshold (say F_{spec}). We call them to be the "trusted" media streams.

We make the following assumptions:

$\mathcal{A}1$. All sensing devices capture the same environment (but optionally, the different aspects of the environment) and provide correlated observations.

$\mathcal{A}2$. The system designer's confidence level in each of the media streams is at least 0.5. This assumption is reasonable since it is not useful to employ a media device which is found to be inaccurate more than half of the time.

$\mathcal{A}3$. The fused probability of the occurrence of event and the overall confidence increase monotonically as the more concurring evidences are obtained based on the streams.

The objective is to determine the confidence level $f_i(t+1)$ of a new non-trusted stream M_i at time instant $t+1$ given that its confidence level at time instant t is $f_i(t)$. In absence of any prior information, $f_i(0) = \epsilon$ (a positive infinitesimal).

4 Proposed Method

The proposed method determines the confidence in a new non-trusted stream using its "Agreement Coefficient" with the trusted stream(s). The agreement coefficient between the two streams is computed based on whether the evidence obtained by the system using them are concurring or contradictory.

4.1 Modelling of the Agreement Coefficient

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

$$\mathit{\Gamma}(t) = \{\gamma_{ik}(t)\} \tag{1}$$

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 system computes 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. Precisely, $\gamma_{ik}(t)$ is computed as:

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

where, $p_i(t) = P(E_t|M_i)$ and $p_k(t) = P(E_t|M_k)$ are the individual probabilities of occurrence of event E 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 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 ($\gamma_{ik} = -1$) [4].

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

$$\gamma_{M_i,\mathbf{M}^{i-1}} = \frac{1}{i-1}\sum_{s=1}^{i-1}\gamma_{si} \tag{3}$$

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 (3) 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 [10]. 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.

4.2 Confidence Fusion

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. Given that the two streams M_i and M_k have their confidence levels f_i and f_k, respectively, the system uses 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:

$$f_{ik} = \frac{f_i \times f_k}{f_i \times f_k + (1 - f_i) \times (1 - f_k)} \tag{4}$$

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 (Refer to assumption $\mathcal{A}2$, in section 3). Second, although the media streams are correlated in their decisions; we assume that they are mutually independent in terms of their confidence levels [4].

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:

$$F_i = \frac{F_{i-1} \times f_i}{F_{i-1} \times f_i + (1 - F_{i-1}) \times (1 - f_i)} \tag{5}$$

4.3 Confidence Building Method

Given a set T of trusted media streams, the confidence level of the media stream M_i is computed as follows:

- Using a voting strategy, we divide the set T of trusted media streams into two subsets T_1 and T_2 (as shown in figure 1). This division is performed based on whether, at the current instant, the evidence obtained by the system using these two subsets are concurring or contradictory. Precisely, the subset, based on which, the system concludes in favor of the occurrence of event E with more than 0.50 probability are put in set T_1 and the rest in set T_2.
- The agreement coefficients γ_{i,T_1} (between the stream M_i and the subsets T_1) and γ_{i,T_2} (between the stream M_i and the subset T_2) are computed as described in section 4.1 (equation 3).
- Next, the system computes the overall confidence F_{T_1} and F_{T_2} in the subsets T_1 and T_2, respectively, (using equation 5).
- Finally, the system computes the confidence $f_i(t+1)$ in the i^{th} stream at time instant $t+1$ as follows:

$$f_i(t+1) = \begin{cases} F_{T_1} \times \frac{f_i(t).e^{\alpha.\gamma_{i,T_1}(t)}}{f_i(t).e^{\alpha.\gamma_{i,T_1}(t)} + (1-f_i(t)).e^{-\alpha.\gamma_{i,T_1}(t)}} & \text{if } F_{T_1} \times \gamma_{i,T_1}(t) \geq \\ & \quad F_{T_2} \times \gamma_{i,T_2} \\ F_{T_2} \times \frac{f_i(t).e^{\alpha.\gamma_{i,T_2}(t)}}{f_i(t).e^{\alpha.\gamma_{i,T_2}(t)} + (1-f_i(t)).e^{-\alpha.\gamma_{i,T_2}(t)}} & \text{otherwise} \end{cases}$$

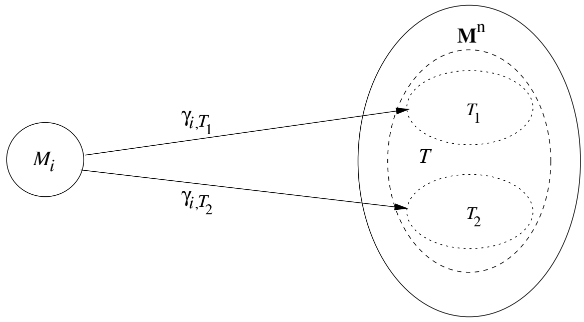


Fig. 1. The agreement coefficient between stream M_i and the subsets T_1 and T_2

The exponential terms $e^{\alpha \cdot \gamma_{i,T_1}(t)}$ and $e^{\alpha \cdot \gamma_{i,T_2}(t)}$ represent the growth in the confidence level at time t. $\alpha \in [0, \infty]$ is used as a growth rate with respect to overall confidence levels F_{T_1} and F_{T_2} of groups T_1 and T_2, respectively. The terms $\gamma_{i,T_1}(t)$ and $\gamma_{i,T_2}(t)$ denote the agreement coefficient at time t between the i^{th} stream and the groups T_1 and T_2, respectively. In the above formulation, the denominator term acts as normalization factor to limit the confidence value within [0,1]. Note that if either T_1 or T_2 is found empty, their fused confidence levels (F_{T_1} and F_{T_2}, respectively) are considered to be of zero value.

5 Experimental Results

We show the utility of our method in a surveillance scenario. The surveillance environment is the corridor of our school building with a system goal to detect events such as humans running, walking and standing in the corridor. We use two video sensors (Canon VC-C50i cameras denoted by V_1 and V_2) to record the video from the two opposite ends of corridor as shown in figure 2. The two cameras are connected to a central PC (Pentium-IV 3.6 GHz). A Picolo-Pro video capture card is used to capture the image data.

For our experiments, we have used data of more than twelve hours which has been recorded using the system consisting of two video cameras. Over the period of more than twelve hours, the noticeable events occurred over for a period

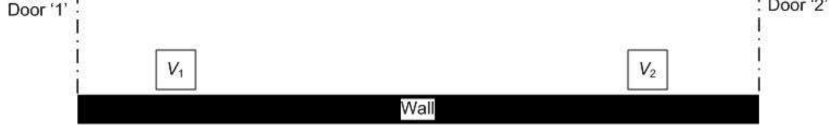


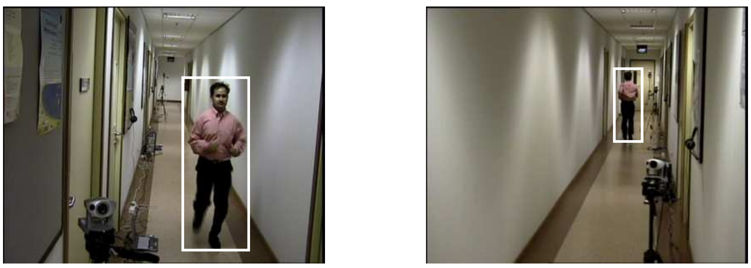
Fig. 2. The layout of the corridor under surveillance

of 1079 seconds. The details of various events and their total durations are as follows - humans standing events for 139 seconds, walking events for 798 seconds and running events for 142 seconds.

The system detects these events by processing the video frames. The video processing involves background modeling and blob detection. The background modeling is performed using an adaptive Gaussian method [11]. For blob detection, the system first segments the foreground from the background using simple 'matching' on the three RGB color channels, and then uses 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. We assume that the blob of an area greater than a threshold corresponds to a human. An example of blob detection (with its bounding rectangle) in a humans "running" event is shown in figure 3. Once the bounding rectangle for each blob is computed, the middle point of the bottom edge of the bounding rectangle is mapped to the actual ground location using the caliberation information of the video cameras. This provides the exact ground location of human in the corridor at a particular time instant.

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. Based on the average distance travelled by human on the ground, a Bayes classifier is first trained and then used to classify an atomic-event to be one of the classes - standing, walking and running.

We present our preliminary results as follows. First, the system performed event detection and classification using only one stream i.e. video stream V_1. By comparing with the ground truth, we found overall accuracy of the video stream V_1 to be 68%. Based on the accuracy, we assigned a confidence level 0.68 to the stream V_1, and designated it to be the "trusted" stream. Note that, in our experiments, the threshold value used for trusted stream is 0.65. Determing the ideal value of this threshold is an issue which we will examine in the future



(a) (b)

Fig. 3. An example: Bounding rectangles along the detected blobs in the video frames of (a) Camera 1 and (b) Camera 2, corresponding to a hummas "running" event

work. Based on the agreement coefficient between the trusted stream V_1 and the other video stream V_2, the system uses our proposed method to compute the confidence in stream V_2.

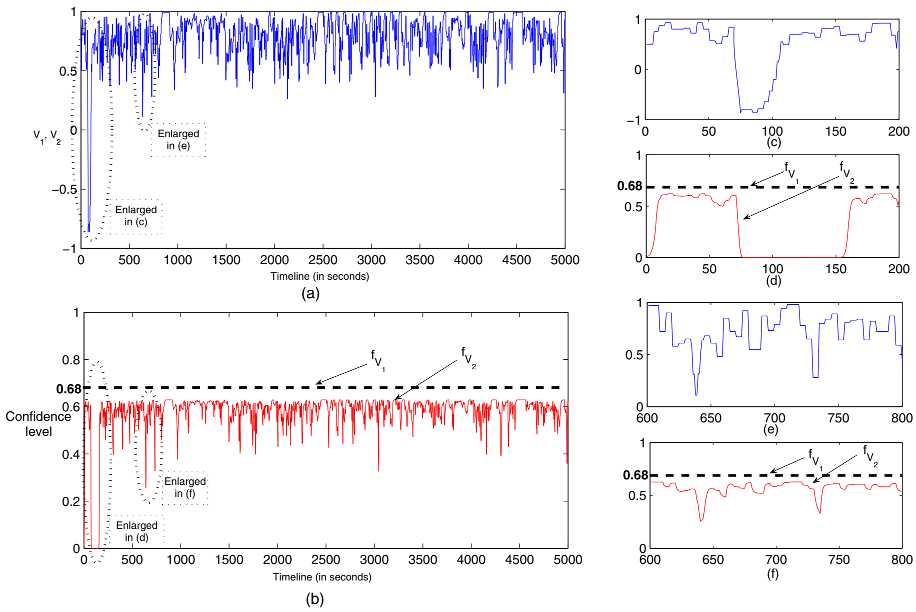


Fig. 4. Confidence building in V_2 stream

Timeline-based confidence building in the stream V_2 is shown in figure 4. Figure 4(a) shows how the agreement coefficient between V_1 and V_2 varies along the timeline, and figure 4(b) depicts how the confidence f_{V_2} in V_2 evolves along the timeline. Figures 4(c)-4(f) show the enlarged portions of some parts of figures 4(a)-4(b). For example, figure 4(c) shows how the agreement coefficient γ_{V_1,V_2} drops down below zero and then consequently figure 4(d) depicts that how confidence also decreases as the agreement coefficient decreases. Once the confidence level drops down at around 75^{th} second along the timeline, the confidence f_{V_2} in stream V_2 also drops to almost zero and it takes approximately another 90 seconds to regain the same confidence level. Similarly, as can be seen in figures 4(e)-4(f), the confidence level f_{V_2} in stream V_2 decreases to 0.25 at time instant 642 and to 0.30 at time instant 740 as the agreement coefficient goes below 0.5, however, in these two cases, the confidence level picks up early close to the confidence level of trusted stream compared to when the agreement coefficient becomes negative.

Note that we have set the value of the growth rate α to be 1. With $\alpha = 1$, the system could gain the confidence level in V_2 upto 0.63. However, with higher growth rate ($\alpha \approx 5$), this maximum achieved confidence can go up to the level of confidence in the trusted stream. Again, determining the ideal value of growth factor α is an issue which is out of scope of this paper and will be investigated in the future.

To verify the utility of our method, we compared the average confidence level of stream V_2 determined using our method with the confidence level which is computed based on its past accuracy. It is observed that both are comparable (0.58 vs 0.60) as shown in Table 1.

Table 1. Comparison of the proposed method with the traditional approach

Method of computing confidence	Confidence level of V_2
Pre-computed confidence	0.60
Our method	0.58 (Average value)

6 Conclusions

This paper proposes a novel method to dynamically compute the confidence levels of new media streams in a multimedia surveillance system. The confidence in a new stream is computed based on the fact whether it provides evidence which concurs or contradicts with the already trusted streams. Though the preliminary results have shown that the confidence level computed using our method is comparable with the confidence level determined based on the traditional approach (past accuracy), we need to investigate in detail how the dynamically varying confidence level can contribute towards more accurate overall results for event detection in multimedia surveillance systems. It will also be interesting to examine the utility of the proposed method for dissimilar sensors such as microphones and motion detectors, and also for the different kinds of events such as audio events - talking, shouting, door knocking and footsteps.

Acknowledgements

We thank Prof Ramesh Jain for suggesting this research problem.

References

1. Atrey, P.K., Maddage, N.C., Kankanhalli, M.S.: Audio based event detection for multimedia surveillance. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2006) V813–816
2. Rama, K.G.S., Atrey, P.K., Singh, V.K., Ramakrishnan, K., Kankanhalli, M.S.: A design methodology for selection and placement of sensors in multimedia surveillance systems. In: The 4th ACM International Workshop on Video Surveillance and Sensor Networks, Santa Barbara, CA, USA (2006)
3. Prati, A., Vezzani, R., Benini, L., Farella, E., Zappi, P.: An integrated multi-modal sensor network for video surveillance. In: The ACM International Workshop on Video Surveillance and Sensor Networks, Singapore (2005) 95–102

4. Atrey, P.K., Kankanhalli, M.S., Jain, R.: A framework for information assimilation in multimedia surveillance systems. ACM Multimedia Systems Journal (2006)
5. Tatbul, N., Buller, M., Hoyt, R., Mullen, S., Zdonik, S.: Confidence-based data management for personal area sensor networks. In: The Workshop on Data Management for Sensor Networks. (2004) 24–31
6. Tavakoli, A., Zhang, J., Son, S.H.: Group-based event detection in undersea sensor networks. In: Second International Workshop on Networked Sensing Systems, San Diego, California, USA (2005)
7. Ioannou, S., Wallace, M., Karpouzis, K., Raouzaiou, A., Kollias, S.: Confidence-based fusion of multiple feature cues for facial expresssion recognition. In: The 14th IEEE International Conference on Fuzzy Systems, Reno, Nevada, USA (2005) 207–212
8. Siegel, M., Wu, H.: Confidence fusion. In: IEEE International Workshop on Robot Sensing. (2004) 96–99
9. Conaire, C.O., Connor, N.O., Cooke, E., Smeaton, A.: Detection thresholding using mutual information. In: International Conference on Computer Vision Theory and Applications, Setubal, Portugal (2006)
10. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys **31(3)** (1999) 264–323
11. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2., Ft. Collins, CO, USA (1999) 252–258