ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 115-122
doi: 10.18178/wcse.2016.06.020

A New Idea of Data Fusion Method Based on Graphic Area for Information System Security Situation Assessment

Yiyang Jia [1 +], Haiyan Wu [2] and Dongxing Jiang [2]

[1] Department of Computer Science, Tsinghua University, Beijing, China
[2] Information Technology Center, Tsinghua University, Beijing, China

Abstract. With the wise used of computer, information security has become more and more concerned by every researcher. In a large number of research directions, information security situation assessment is a hot topic. On current situation, researchers have made a lot of effort about in the field of model and framework. In fact, there are some proposed models or frameworks have showed a good result, and do withstand the test of long term practice. However, as we focus our attention on the study of models and frameworks, we ignore one of the most important points for information security situation assessment which is data fusion. Even we have a mature model of how to organize these data, we still need an effective data fusion method to fuse them, and evaluate the information security situation according to the fuse result. Some researchers have conducted a series of studies on this, but their studies are more concerned about data fuse architecture rather than the method itself. This paper proposes a new idea of data fusion method based on graphic area, we use graphs to represent the data to be fused, and use the area of graph to represent the fusion result. The proposed method is not only fuses data more accuracy but can display the result in a more intuitive way. With this method, the fusion result is easier to understand and has a better visualization. We use data from Tsinghua University information system to verify the proposed method. And the comparison result shows that our method can fuse data more comprehensively and shows a better visualization effects.

Keywords: information system, security assessment, data fusion, hierarchical model.

1. Introduction

With the rapid development of computer networks, more and more people begin to pay attention to security issues. As one of the most important component in the security field, security assessment has become a debated direction among researchers. The concept of security assessment was put forward by Endsley [1] in 1988 and it was first used in the military field. With the booming of the internet, it was gradually applied to computer and networks. Security assessment means that we obtain information from various sorts of sensors on the edge of the system, then fuse and evaluate them, thus we can assess the security situational of the system [2], [3]. After a long period of studies, information security situation assessment has formed a classic analysis model of three levels [4]. Figure 1 has shown the model.

The first level is situation understanding, the second level is situation evaluation and the third level is situation forecast. Situation understanding can also be divided into two steps: factors acquisition and situation understanding. Factors acquisition is the premise of the security assessment, and it means obtain information which would impact system from various sensors. Situation understanding means dealing with the factor and analyzing them, this is the basic of assessment. The second level, situation evaluation refers to analyzing and fusing data quantitatively in order to get security status and weak link of current system. It is also the core of information security situation assessment.

[+] Corresponding author. Tel.: + 86 18810914043.
E-mail address: jyy13@mails.tsinghua.edu.cn.

As the core of information security situation assessment, how to evaluate the factors and how to fuse them play a key role. We call this part of the work data fusion [5], [6]. There are many researchers has studied in this field. Hervaldo S. Carvalho [7] has proposed a general data fusion architecture based on UML and using a taxonomy based on the definitions of raw data and variables or tasks. Ronald R. Yager [8] has provided a general framework for data fusion based on a voting like process that tries to adjudicate conflict among the data. David L. Hall [9] gives a comprehensive introduction to multi-sensor data fusion. The existing research always concerned about framework or architecture. However, data fusion method itself has not received much attention, especially for the research of data fusion method in the hierarchical evaluation model.

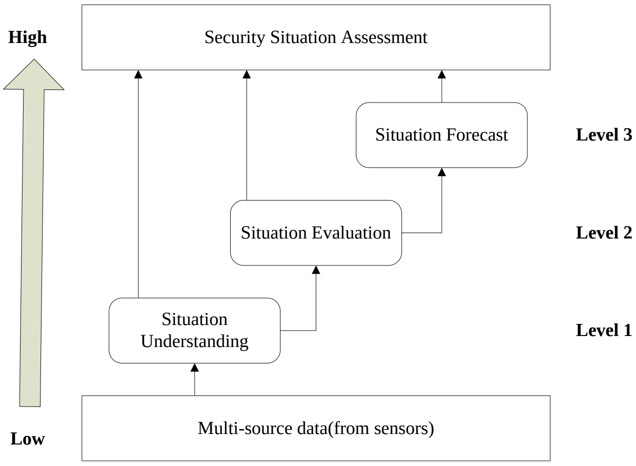

Fig. 1: Hierarchical model of security situation assessment.

In this paper, we propose a data fusion method based on graphic area, and we can achieve a better effect by fusing multi-sensor data in this method. Compared with other methods, this method has less computation work and is more convenient to implement, and it is more suitable for the hierarchical security assessment model. This method can not only fusing data into a more representative value but also has a better visual effect and makes users understand the weak link of system more intuitively, therefore, they can evaluate the security situation more accuracy.

The reminder of this paper is organized as follows. Section 2 will elaborate the method this paper proposed. Experiment data and result analysis will show in Section 3. At last, we will give conclusion in Section 4.

2. Proposed Method

As we have mentioned above, our method is proposed for information security situation assessment, especially for hierarchical model in this field. So we will first introduce the assessment model and scenarios in this section. Then we will introduce data preprocessing. At last, we will give a detailed elaboration about our method proposed in this paper.

2.1. Situation Assessment Model and Index System

In this paper, we use our method in the information security situation assessment. And we use a hierarchical model as Figure 2 to describe this problem.

As the figure shows, we divided information system into 3 levels: sub-system level, component level, index level. Our data fusion method is used in the integration of data on each level. For example, if we want to evaluate the security situation of data base on level 2, we need to use our method to fuse the index data of data base on level 3.

The first to carry out in the whole assessment process is index fusion on the bottom level. Obviously, we cannot get these index data directly, because they are fused from raw data. The so-called raw data refers to the data obtained directly from a variety of sensors. How to choose these raw data so that they can reflect the situation of system is another problem to be solved. Therefore, we design a two-layer index system which shows in the table below.

From the table, we can see that each index on level 3 has a set of secondary indexes. And all of these secondary indexes can be accessed from variety of sensors. They all constitute our raw data, and the whole assessment work is based on these data.

Major Index	**Secondary Index**
Vulnerability Index	Number of vulnerabilities, Severity of vulnerabilities, Exploitability of vulnerabilities, Etc.
Environment Index	Number of system views, Memory usage, Current flow information, Response time of system and etc.
Importance Index	Importance of data, Recoverability of data and etc.
Anomaly Index	Number of invasion, Severity of invasion, Duration of invasion, Source number of invasion and etc.

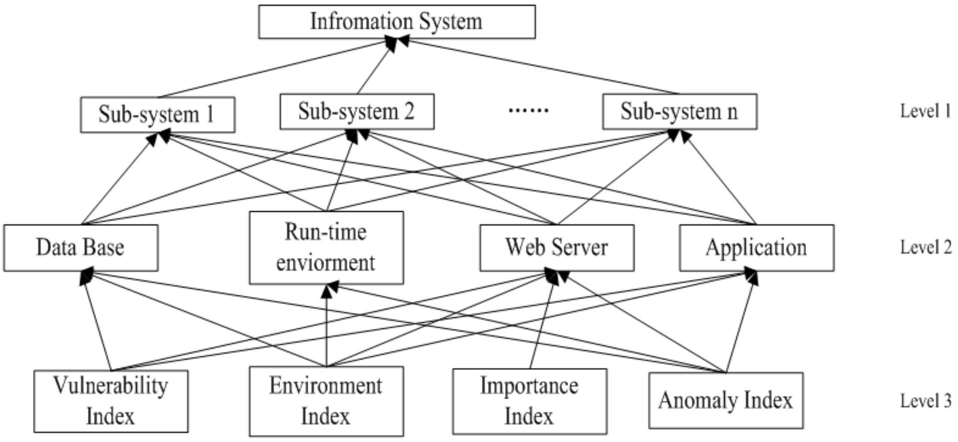

Fig. 2: Hierarchical model of information system.

2.2. Data Pre-processing

Different sensor leads to different data. Since we have various sensors, the raw data can be very different, whether the scope or the way it is described. That means we need to preprocess these data before we use them. Figure 3 shows the process of data preprocessing.

After the data preprocessing, we can get quantitative index data, and then fuse these index data.

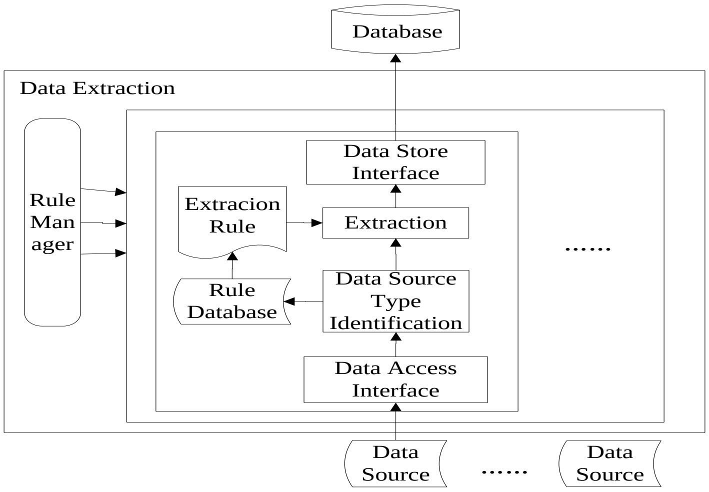

Fig. 3: Architecture of data pre-process.

2.3. Data Fusion Method Based on Graphic Area

In this part, we will introduce the data fusion method based on graphic area in detail. When evaluating the security situation of the information system, we want to know the current security situation of the system at every moment. So when we evaluate the system, we collect all kinds of data at the same time as a sequence. According to the hierarchical model we have mentioned before, it can be seen that in order to evaluate the security situation of system on the top level, we need to fuse data start from the bottom level. In order to describe our method clearly, we give some basic definitions which will be used later first.

Definition. We use *S* to represent the security situational value of the entire information system, which is our target. S_i represents the security situational value of each sub-system; *i* represent the id of sub-system, $\boldsymbol{i} \in [\mathbf{1}, \boldsymbol{n}]$; *n* represent the number of sub-systems. Similarly, we use C_j to represent the security situation of component in each sub-system. *j* means the id of the 4 components, $\boldsymbol{j} \in [\mathbf{1}, \mathbf{4}]$. w_i is the weight of each sub-system, it is assigned according to the influence degree of each sub-system to the entire system, so is u_i which means the weight of each component to sub-system. $\vec{\boldsymbol{I}}$ means the major index vector, and $\vec{\boldsymbol{v}}$ means the integration vector of the credibility and importance of each major index which will be given by expert. Based on these, we give the function of *S* as bellow:

$$S = \sum_{i=1}^{n} S_i \times w_i \qquad S_i = \sum_{j=1}^{4} C_j \times u_j \qquad C_j = \vec{v} \cdot \bar{I}$$

Among them, *S* can be obtained by S_i, and S_i can be obtained by C_j. C_j can be obtained by major index in index system, so how to get quantitative major index $\bar{I}$ is the focal point of this section which is also the key of this paper.

In this method, we use data of each time point as a fuse unit. These data constitute the sequences to be fused. And we will pick up two sequences from history data. These two sequences represent the best situation and the worst situation. We use these two sequences as a normalized standard. Next, we use a multidimensional coordinate system on a plane to fuse the data. Specific steps are as follows:

Step1. Get the data after the preprocess, and record the data as a sequence such as *D (d1, d2,…, dn)*. Take environment index as an example, d_i means the secondary index of environment index.

Step2. Pick up two reference sequences. According to the physical means of secondary indexes, define a good sequence *g* and a bad sequence *b*, the value of each secondary index in *g* is the best in history data, and the value of *b* is the worst in history data.

Step3. Data normalization. We use the reference sequences as standard to normalized index data. The normalization formula is defined as bellow:

$$d_i(k) = \frac{d_i(k) - min}{max - min}$$

The *k* means the time point of the data. We will fuse multiple set of data at one time, so *k* is the identification of time points.

After the first three steps, we have got the normalized data, and next step is to fuse them in a coordinate system.

Step4. First of this step, we will construct an n-dimension plane coordinate system. *n* refers to the number of dimensions in the sequence we are going to evaluate. For example, when we evaluate the environment index, we need to take access number, highest response time, lowest response time, average response time and the failure rate these 5 secondary indexes into consideration. So we construct a 5-dimension plane coordinate system which means *n* equals 5. Whether the whole plane can be divided equally between each dimension can be set through a weight vector. Coordinate system will be divided into n parts, the size of each part can be adjusted according to the influence of this factor to the whole system. The greater impact of the factors, the larger the size is, and vice versa. Because our fusion is based on area, the larger part will impact the result more in the calculation process. We define the weight vector as $\vec{\omega}(\omega_1, \omega_2, \dots, \omega_n)$, vector $\vec{\omega}$ satisfies the following condition. Using this weight vector, we can get the arc size of each factor in the plane coordinate system. We record the arc size as $\vec{r}$, and use the following formula to calculate it:

$$\sum_{i=1}^{n} \omega_i = 1 \qquad \vec{r} = 2\pi \cdot \vec{\omega}$$

Now, we can get an n-dimension plane coordinate system, each axe in this coordinate system represents one dimension data to be fused, the coordinate system may be look like as Figure 4.

Step5. Calculate the graphic area on the coordinate system. We use the measure of the area to represents the security situation of the system. The measure of the area is calculated via the formula bellow:

$$SA = \frac{1}{2}\sum_{i=1}^{n-1} d_i \cdot \sin r_i \cdot d_{i+1} + \frac{1}{2} d_n \cdot \sin r_n \cdot d_1$$

Step6. Because of *d* is in the range from 0 to 1, the range of the calculated *SA* must be 0 to $\boldsymbol{\pi}$. In order to show the security situation assessment result more intuitively, we choose to normalize the calculated *SA* so that the result can be range from 0 to 1.

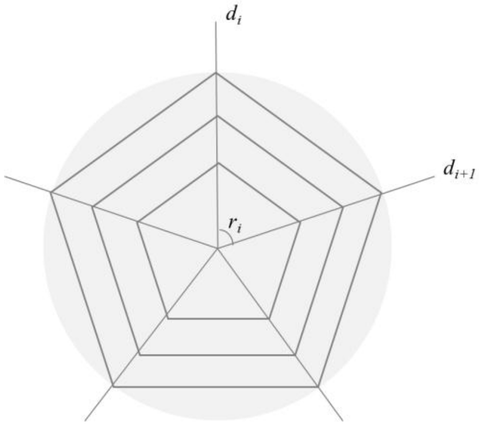

Fig. 4: Coordinate system of indexes.

2.4. Visual Expression

This method has a significant advantage compared to other methods. This advantage is having a better visual effect. We can show the coordinate system to users directly, the intuitive coordinate system allows users to understand the security situation in all aspects, not just feeling by an abstract value. Security situation at different time points can be displayed in the same coordinate system by setting area with different color and transparency. This approach enables users to grasp the trends of security situation of the system, and thus to improve the weakness more purposeful.

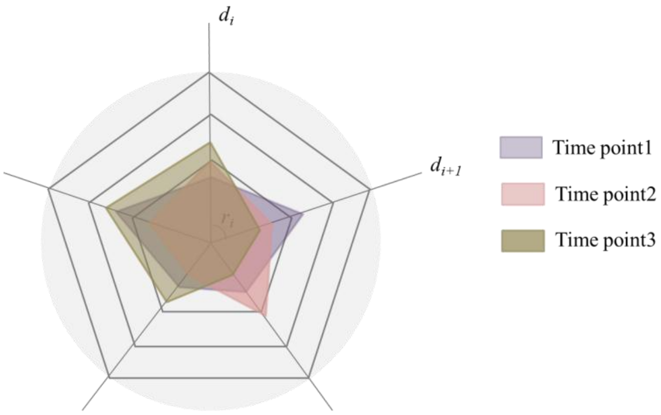

Fig. 5: Comparison of index on different time points.

In addition to compare the performance of various aspects of the system at different time points, we can also make a comparison of different components. In our hierarchical model, different components may have same indexes. We can also make a comparative analysis of the various components in this dimension, so as to understand the weak points of the system. The comparison coordinate should be shown like the Figure 5.

3. Experiment Analysis

In order to prove that our method can be applied to information system security situation assessment, we implement a series experiment use our information system in school. We use the *Learning Class* information system as our experiment system, and pick one week data as our experiment data to implement.

Learning Class is one of our sub-systems in our information system. This sub-system has two web servers, one database and an application server. All of these components form the second level in our hierarchical model. And we will use our method to fuse the index data of each component, and calculate the system security situation according to the formula in Section 3.

Secondary Index	**Nov. 21th**	**Nov. 22th**	**Nov. 23th**	**Nov. 24th**	**Nov. 25th**	**Nov. 26th**	**Nov. 27th**
Access Number	169312	140502	208216	235804	250197	229972	131983
Highest Resp Time	2602.35	3837.69	2738.82	2703.14	3558.69	2256.65	2504.09
Lowest Resp Time	415.81	566.06	407.36	296.57	360.39	325.14	298.86
Average Resp Time	1361.28	1790.75	1341.48	1162.88	1473.2	1067.42	1157.01
Failure Rate	0.395%	1.008%	0.349%	0.351%	0.07%	0.699%	0.348%

Table II lists the secondary index of environment index from November 21th to November 27th, and we will use these data to carry out the experiment. And we set the best sequence g and the worst sequence b as bellow:

$$g = (100, 500, 1000, 100000, 0) \qquad b = (1000, 2500, 5000, 400000, 5)$$

The data after normalized is shown in the matrix:

$$\begin{bmatrix} 0.3509 & 0.4306 & 0.4006 & 0.2310 & 0.0791 \\ 0.5178 & 0.6454 & 0.7094 & 0.1350 & 0.2016 \\ 0.3415 & 0.4207 & 0.4347 & 0.3607 & 0.0697 \\ 0.2184 & 0.3314 & 0.4258 & 0.4527 & 0.0702 \\ 0.2893 & 0.4866 & 0.6397 & 0.5006 & 0.0139 \\ 0.2502 & 0.2837 & 0.3142 & 0.4332 & 0.1397 \\ 0.2209 & 0.3285 & 0.3760 & 0.1066 & 0.0696 \end{bmatrix}$$

In the matrix, rows represent the data of one day (seven days in total), and columns represent the secondary index of each day, and in environment index are 5 secondary indexes. In order to simplify the calculation, we set vector $\vec{\omega}$ as *1/n*, which is *1/5* in this case. According to the method we proposed in Section 3, we calculate the *SA* of this index.

$$SA= (0.2198, 0.4847, 0.2531, 0.2156, 0.3725, 0.1863, 0.1232)$$

In this case, the worse the index is, the greater the value is. So the bigger the *SA* we calculated means the worse the security situation. In order to meet the normal logical thinking, we normalized the *SA* again, and get the final security situation of the environment index as bellow:

$$S(Environment\ Index) = (0.7802, 0.5153, 0.7469, 0.7844, 0.6275, 0.8137, 0.8768)$$

We also use another two methods to fuse the same data and compare them with our method. The result is shown in Figure 6.

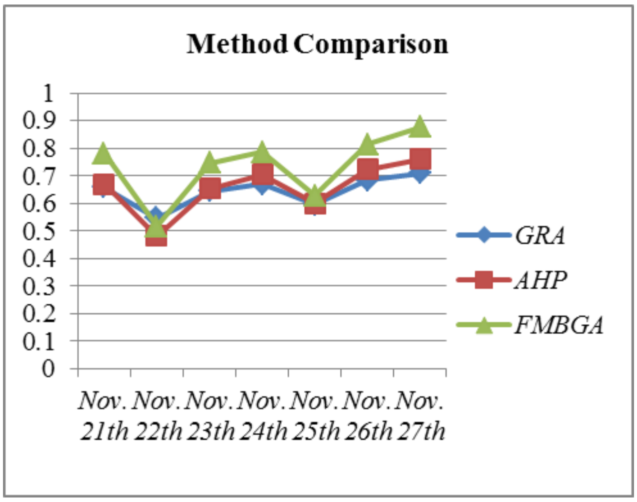

Fig. 6: Method comparison of environment index.

In the Figure 6, our method is the green one, we call it FMBGA (fusion method based on graphic area). The blue one is grey relational analysis method, this method analyzes the weight of each secondary index and fuse them by the weight. The red one is classical AHP method which just normalizes data and weight data simply. From the result, we can see that these three methods all can reflect the security situational of the index, but our method is able to reflect the trend of change more obvious, and our method can be displayed more intuitively. The displayed form will be introduced later.

We also calculate the vulnerability index and anomaly index as Figure 7 and Figure 8. We use the same approach to calculate the security situational value on each level, and set the **w** and **u** to be average. The final security situational value of this week is:

$$S = (0.6648, 0.4647, 0.6534, 0.7327, 0.6252, 0.737, 0.7783)$$

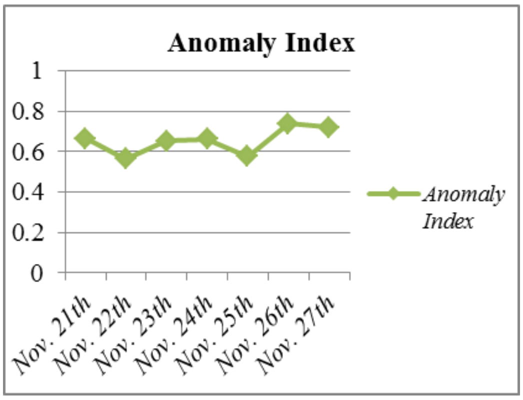

Fig. 7: Security situation of anomaly index.

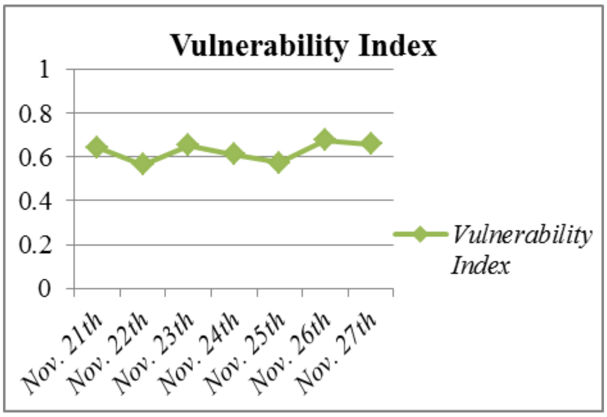

Fig. 8: Security situation of vulnerability index.

In addition to being able to view the security situation of the system through the final value, we also provide a more intuitive way to check the situation which we have mentioned before, used coordinate system. For instance, if we want to know the information of the database in one day, we can just see through the coordinate shown in the Figure 9. At the same time, we can also compare the results in different time points. Comparison of different view can make administrators grasp the situation of the system better, which is not provided by any other methods.

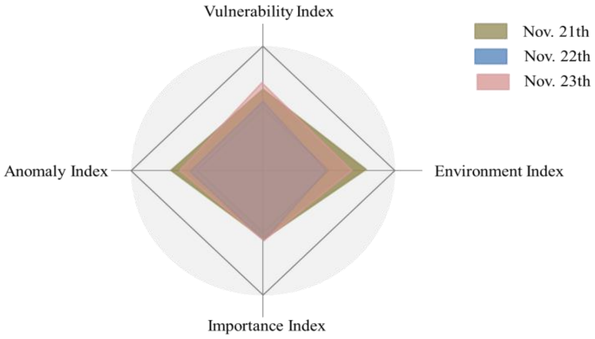

Fig. 9. Comparison of web server for 3 days.

4. Conclusion

In this paper, we provide a new idea of data fusion method for information security situation assessment, which is based on graphic area. We also introduce our hierarchical model of information security situation assessment and the 2-layer index system. In the experiment, we use the data for a week come from *Learning Class* information system. We compare our data fusion method with another two fusion methods, and through the result of experiment, we can see that our method can reflect the trend of change more obvious. And the representation of our fusion method is more intuitive to users.

This method provides a complete process from data collection to data fusion. We calculate a quantitative value range from 0 to 1 to represent the security situation, the higher the value, the more stable the system. In our method, we use a graphic area to fuse the data in different dimension, and evaluate the security situation via this. With this method, the result can be shown in a coordinate system which is more humane and more intuitive for users.

Though we have proposed a new idea for data fusion, and have verified that this idea is applicable to security information in our hierarchical model, it still needs more verification in other fields and problems. And this will be our next direction.

5. References

[1] Endsley, Mica R., and Daniel J. Garland, eds. *Situation awareness analysis and measurement*. CRC Press, 2000.

[2] Endsley M R. Design and evaluation for situation awareness enhancement[C]*//Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* SAGE Publications, 1988, 32(2): 97-101.

[3] Endsley M R. Design and evaluation for situation awareness enhancement[C]*//Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* SAGE Publications, 1988, 32(2): 97-101.

[4] Salerno, John, Michael Hinman, and Douglas Boulware. *Building a framework for situation awareness*. AIR FORCE RESEARCH LAB ROME NY INFORMATION DIRECTORATE, 2004.

[5] Bass T. Intrusion detection systems and multisensor data fusion[J]. *Communications of the ACM*, 2000, 43(4): 99-105.

[6] Waltz, Edward, and James Llinas. Multisensor data fusion. Vol. 685. Norwood, MA: Artech house, 1990.

[7] Carvalho H S, Heinzelman W B, Murphy A L, et al. A general data fusion architecture[C]// Information Fusion, 2003. Proceedings of the Sixth International Conference of. IEEE, 2003:1465-1472.

[8] Yager R R. A framework for multi-source data fusion[J]. *Information Sciences*, 2004, 163(1-3):175-200.

[9] Hall D L, Llinas J. Introduction to multi-sensor data fusion[C]// Proceedings - IEEE International Symposium on Circuits and Systems. 1998:537-540 vol.6.