World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:12, 2015

# Multimedia Data Fusion for Event Detection in Twitter by Using Dempster-Shafer Evidence Theory

Samar M. Alqhtani, Suhuai Luo, Brian Regan



***Abstract*—**Data fusion technology can be the best way to extract useful information from multiple sources of data. It has been widely applied in various applications. This paper presents a data fusion approach in multimedia data for event detection in twitter by using Dempster-Shafer evidence theory. The methodology applies a mining algorithm to detect the event. There are two types of data in the fusion. The first is features extracted from text by using the bag-of-words method which is calculated using the term frequency-inverse document frequency (TF-IDF). The second is the visual features extracted by applying scale-invariant feature transform (SIFT). The Dempster - Shafer theory of evidence is applied in order to fuse the information from these two sources. Our experiments have indicated that comparing to the approaches using individual data source, the proposed data fusion approach can increase the prediction accuracy for event detection. The experimental result showed that the proposed method achieved a high accuracy of 0.97, comparing with 0.93 with texts only, and 0.86 with images only.

***Keywords*—**Data fusion, Dempster-Shafer theory, data mining, event detection.

## I. Introduction

MULTIMEDIA data fusion is the manner in which the different features of multimedia is combined with an aim of analyzing specific media tasks. The process can also be regarded as multimodal fusion. However, to obtain a good understanding of the data, multimedia analysis of this multimodal data has to take place. The most common examples of multimedia analysis are semantic concept detection, audio-visual speaker detection, and human tracking and event detection. In such cases, the multimedia data used can be either sensory or non-sensory. Examples of sensory multimedia task are audio, video or RFID while the non-sensory are like the online resources such as database and WWW resources [1].

The aim of fusion is to improve on the quality, therefore multimedia analysis involves fusion of the available modalities to ensure the output has a better accuracy and the decision making process is reliable. A good example is the use of audio features together with the visual features plus text input while analyzing a sporting event represented in a video. It is however important to note that fusion will increase the cost and make the system analysis more complex [2]. However, it is good to note that:

i. The media can vary in format and rates hence a video can be captured at a rate different from the audio;
ii. The media streams have different processing times hence the chosen strategy has to consider this; iii) the media modalities are either correlated or independent, and the modalities vary in the confidence level required to finish the task; and
iii. The fusion process must take into consideration some cost that is required for capturing and processing of the media.

Samar M. Alqhtani, Suhuai Luo and Brian Regan are with the School of Design, Communication and IT, the University of Newcastle, Callaghan NSW 2308, Australia (e-mail: Samar.alqhtani@uon.edu.au, suhuai.luo@newcastle.edu.au, brian.regan@newcastle.edu.au).

Multimodal fusion is defined as the combination of several multimedia sources plus their features to complement the analysis of the performance. The levels of multimodal fusion can be classified into three, namely: feature level (early fusion) decision level (late fusion) and the combination of the two which is referred as hybrid fusion.

### *A. Feature Level Multimodal Fusion*

Feature level fusion is referred to as early level multimodal fusion and involves the picking of the ideal features from input data. The features are combined and the outcome is forwarded to a single analysis unit (AU) to carry out the analysis. The media stream has distinct features with varying properties. A good example is the feature fusion which combines multimodal features like the skin color and motion cues. Therefore, the combination of the features received is combined into a single sematic level decision [3].

Several features exist and can be combined to create the desired outcome. Examples of possible features are: i) visual features which can be based on color, texture and shape; ii) text features which are possible to be extracted from ASR, OCR, video closed caption text and possible production metadata; iii) audio features which are normally generated according to their FFT or MFCC coupled with features such as ZCR, LPC, volume standard deviation, non- silence ratio and pitch; iv) motion features which are frequently represented as kinetic energy form, hence giving the possibility of measuring the pixel fluctuation in relation to shot, motion direction, magnitude histogram, optical flows and motion pattern formation direction; and v) metadata which is used to complement the data during the production process. Examples are the time stamp, name, image source (video), and finally shots locations [4].

### *B. Decision Level Multimodal Fusion*

It is sometimes called late fusion approach and its analysis unit normally first provides the system local decision $D_1$ to $D_n$ that are normally obtained based on individual features $F_1$ to $F_n$. Using decision fusion ($DF$) the system can be combined DF unit to result to a fused decision vector and that may be analyzed further and further to obtain the final decision output $D$ regarding the task or possible analysis [5].

### C. Hybrid Level Multimodal Fusion

Hybrid level multimodal fusion is meant at to combine both the advantages accrued from the Decision level multimodal and Feature level multimodal fusion. The features in this case are in the first instance fused with a *FF* unit and then the resultant vector is analyzed by an *AU*. Consequently, the individual features are studied under other completely different *AU*s together with other decision features using the DF units [6]. Further fusion occurs in the latter stages of all the decision obtained as the final decision.

## II. Fusion of Text and Image in Social Media

Increased use of social media like Twitter, Facebook, and Instagram has increased the volume of the flowing data to deal with in terms of analyzing and data extraction. These networks have gained huge acceptance and have become part and parcel of the daily lives of so many individuals. As a result, most of these network sites contain a full of significant volume of multimedia data waiting to be mined, as well as analyzed. Social networks are full of different types of content such as multimedia including images, and texts.

Isson and Harriot noted that text mining helps social media analytics since the technology can help sift media-generating text into logical clusters or categories that can be assessed qualitatively against quantitative business metrics [7]. The capacity to use text mining algorithms efficiently when it comes to text and image data is important for a wide range of applications. Social network sites require text mining algorithms for an extensive range of applications such as clustering, classification and keyword search. While classification and search as recognized application for a wide range of situations, social networks have a better structure in terms of links and text. On the other hand, image mining helps to make associations between different images in social media sites as they have large image databases [8]. Mining images necessitates the extraction of the main features of the images regarding particular criteria. After extraction, the image descriptions and feature vectors are submitted to the mining process.

The potential of text mining, image mining by content, and fusion text and image mining in social media affords a real opportunity for supporting innovation and development of new knowledge which is widely used in a wide range of areas such as business and competitive intelligence, and national security among others. They enable individuals and organizations to make sense of the vast data resources and information and to leverage value [7]. Therefore, fusion is applied for text and image by mere combination of the image and the text features [9]. Moreover, fusing image documents and text documents makes it possible to improve image clusters in social media [10]. It is proper to note that in a fusion method, where the text mining score is dismal in comparison to the threshold, the text mining in such as case cannot be depended upon, hence the tweet is solely classified using the image only and vice versa [11].

Twitter has created a platform where people share among other things real life events happening in real time. However, considering that most of the tweets are meaningless, there is need to design a mechanism that detects crucial shared events in almost real time [12]. Several events that happen and are tweeted about, such as concerts, disaster, sports events, public celebrations, or even protests should be directly detected by such software's. However, these events can be presented online in terms of text data, image data or both hence the technology should be able to draw out the difference and notice all cases [13].

## III. Dempster-Shafer Data Fusion Theory

Dempster-Shafer theory (DST) is more focused on the belief, unlike the Bayes theory which focuses on the probability and DST is widely used in classification problems. Dempster-Shafer evidence theory offers an alternative to traditional probabilistic theory for the mathematical representation of uncertainty [14]. Dempster-Shafer evidence theory enjoys the advantage that it has the ability to deal luck of ideal information (ignorance) and missing details in the data. The second advantage is its ability to deal with union of classes. Dempster-Shafer data fusion theory is a mathematical theory of evidence normally used in a situation observations from varying sources are summed together to give a degree of belief that considers all the evidence presented [15]. The theory of evidence assigns a belief mass to each element of the power set. Formally, a function $m: 2^{\Omega} \rightarrow [0,1]$ is called a basic belief assignment (BBA) and represented by the mass function m, when the subsets of Ω with non-zero mass assignment are called focal element, and it has two properties. First, the mass of the empty set is zero: $m(\phi) = 0$, and the second is the masses of the remaining members of the power set add up to a total of 1:

$$\sum_{A \subset \Omega} m(A) = 1$$

where $\phi$ denotes the null set, and $m(A)$ is called the basic belief assignment of A, where A is a subset of Ω.

The initial requirement of the Dempster-Shafer theory is mass dependent. It will require that masses be assigned to it meaningfully in different ways. At the same time, Dempster-Shafer theory will require preliminary prior information that is present at that particular time and the masses should be assigned in such a way that it shows the knowledge of the system [16]. The principle of operation is based on the knowledge that the level of belief for a given question is obtainable from other subjective probabilities of the other relevant question. Dempster-Shafer's rule is therefore used to combine the varying degrees of belief in case independent evidence is available [17].

The aim of the theory is to decompose the evidence so that probability judgment is separately based on each component of evidence which is to be combined by the Dempster's rule. Therefore, the rule combines parallel belief functions that maybe unrelated to create a pool of belief function that is a

summation of their features.

Dempster-Shafer theory contains two new theories that are missing in Bayes theory. These two theories are notions of support and plausibility. In case the support for the target becomes "quick" it is defined to be the total mass of all the states referring to it as the "fast".

$$spt(A) = \sum_{B|B \subseteq A} m(B)$$

where *spt (A),* is defined as the total mass *m* of all states implying the "*A*" state.

The support is an example of a loose but lower limit to the uncertainty. On the other hand, a loose upper unit to the uncertainty is the plausibility. The definition states that even for the fast state, the total mass of all other states will not contradict the fast state [16].

$$pls(A) = \sum_{B|A \cap B = \varnothing} m(B)$$

where *pls(A),* is defined as the total mass *m* of all states that doesn't contradict the "*A*" state.

Data fusion is a relatively new field with most methods still regarded as unreliable. However, Demister-Shafer theory though relatively new is more reliable compared to the rest in data fusion where it is applied in twitter data before for location estimation problem for the events detected [18].

## IV. The Proposed Method

In this section, we explain the detail of each step of the proposed system. Before that, we monitor a Twitter stream to pick up tweets having both text and image, and store them into a database. Then, we detect the event in text data only, image data only, and fuse the image with the text in the last stage of the method.

### *A. Text Data*

Text data mining is useful for research into social media because it gives researchers the ability to automatically detect events in Twitter. We use the text data to detect event in this step. Tweet messages are written in sentences for in general of which the maximum number of letters is 140. To do event detection by using text data in Twitter, we filtered out tweets that contain non-Latin characters, trying to maintain a corpus of English tweets. Although we managed to remove all East Asian tweets, our corpus still contained some non–English tweets mainly in Spanish and Dutch. We converted all words to lowercase in the tweets. Then we follow the procedure: first, tokenize by convert the string to a list of tokens based on whitespace. This process also removes punctuation marks from the text. Second, filter our text data by use of two types: i) stop word filtering which eliminates the words which are common and their presence does not tell us anything about the dataset, such as: the, and, for, etc. and ii) stem filtering which reduces each word to its stem, removing any prefixes or suffixes. Finally, indexing the data after filtering by using TF-IDF which is a weighting scheme that weighs features in tweets based on how often the word occurs in an individual tweet compared with how often it occurs in other tweets [19]. The term weighting is a key technique in information retrieval (IR) and we explore its use in visual-word feature representation. Then by apply the popular term weighting schemes in IR, we achieve the word feature vectors.

### *B. Image Data*

Our method for image data achieves efficiency through careful feature selection by using principal component analysis (PCA) [20]. Before that we extracted our visual features by using scale-invariant feature transform (SIFT) to automatically detect keypoints from images [21]. Then we use techniques which cluster the keypoint descriptors in their feature space into a large number of clusters using the K-means clustering algorithm and encode each keypoint by the index of the cluster to which it belongs that is called the vector quantization (VQ) [22]. Then we consider each cluster as a visual word that represents a specific pattern shared by the keypoints in that cluster. Therefore, the clustering process generates a visual-word vocabulary describing different patterns in images. The number of clusters determines the size of the vocabulary. By mapping the keypoints to visual words, we can represent each image as a "bag of visual words". Finally, the bag-of-visual-words representation can be converted into a visual-word vector similar to the term vector of words.

### *C. Dempster-Shafer Fusion*

Data fusion is a relatively new field with most methods still regarded unreliable. However, Demister-Shafer theory though relatively new is more reliable compared to the rest in data fusion. Dempster-Shafer theory requires some preliminary assignment of masses that reflects our initial knowledge of the system, including the "unknown" state. The key concept is basic probability assignment or mass assignment. The method is depicted in Fig. 1.

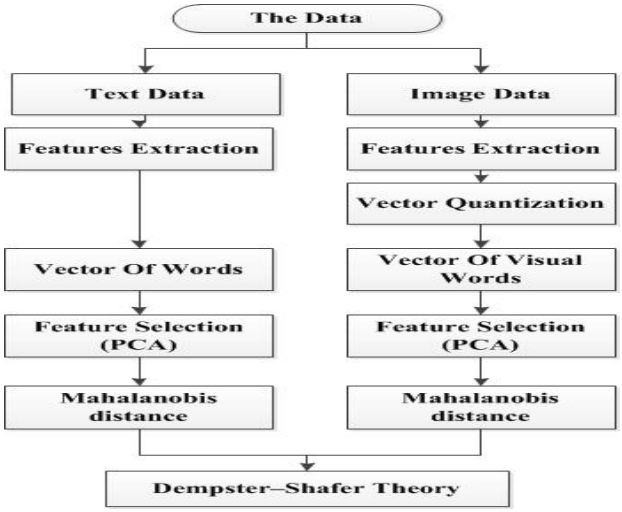


Fig. 1 The block diagram of fusion method

We consider two classes in the Dempster-Shafer theory of

evidence. A feature belongs for either text *t* or image $\bar{t}$ , and $\Theta$ refers to uncertainty inherent in the theory of evidence. All this constitute the frame of discernment Θ in our case:

$$\Theta = \{t, \bar{t}, \theta\}$$

For each feature, we calculate evidence for each class with condition:

$$\mu_i(t) + \mu_i(\bar{t}) + \mu_i(\theta) = 1 \quad (1)$$

where *i* is the number of evidence and $\mu$ obtained for each order statistic. The Mahalanobis distance d is calculated by using:

$$d = (m_t - x)' \sum_t^{-1} (m_t - x) \quad (2)$$

where *x* represents the feature, and ( $m_t$, $\sum_t$ ) are the mean and covariance matrix of the training set. After that, the maximum $d_{max}$ and minimum $d_{min}$ values are obtained for normalising, and the complement to one is computed:

$$d' = 1 - \frac{d - d_{\min}}{d_{\max} - d_{\min}} \quad (3)$$

The standard deviation for the distance's values for all the features obtained in (3) is taken as the uncertainty $\mu_1(\theta) = \sigma$. In order to obtain $\mu_1(t)$ we use the condition in (1). The result is:

$$\mu_1(t) = d'(1 - \sigma) \quad (4)$$

The results we obtain from (4) give us the new evidence masses, and the results belong to the training set. On the other hand, the values for not belonging to the training set are given as:

$$\mu_1(\bar{t}) = 1 - \mu_1(t) - \mu_1(\theta) = (1 - d')(1 - \sigma) \quad (5)$$

Dempster-Shafer theory gives a rule for calculating the confidence measure of each state, based on data from different evidences. Dempster's rule of combination has been used as a fusion strategy to fuse different kinds of data, as given in (6) which fuse two types of data:

$$m^{1,2}(f) = \frac{\sum_{t \cap \bar{t} = f, f \neq \phi} m^1(t) m^2(\bar{t})}{1 - k} \quad (6)$$

where $k = \sum_{t \cap \bar{t} = \phi} m^1(t) m^2(\bar{t})$, and k stands for the basic probability mass associated with conflict, which is determined by summarizing the products of the basic belief assignments (BBA's) of all sets where the intersection is null. *f* is the intersection of states t and $\bar{t}$ in (5); $m^{1,2}(f)$ is the new evidence updated by the evidence sources $m^1(t)$ from sensor 1 (text) and $m^2(\bar{t})$ from sensor 2 (image).

## V. Experiment and Result

In the experiment, the data extracted from twitter which contains texts and photos posted about the Napa Earthquake 2014, California, and it's collected from the Twitter stream from 25 August 2014 to 30 August 2014. We train our algorithms on our data. We divided the data into three equal parts. We used the earliest two thirds of the data as training and validation sets.

We prepared three groups of features for each tweet to detect the event as:

- The First group: the text features extracted for text mining.
- The Second group: the image features extracted for image mining.
- The Third group: the features fusion by Dempster-Shafer theory for text and image features.

Lastly, we measure the accuracy for each data type by applying the equation:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

where A represents the accuracy for the event detection method, TP, TN, FP and FN represents true positive, true negative, false positive and false negative respectively. In our classification, earthquake is happened class is a true positive.

From our experiment, we achieve that accuracy from event detection model for the Dempster-Shafer fusion of text and image gave more accurate result and made the event detection more effective. The result is shown in Table I.

TABLE I
The Accuracy's Result for Each Method

| The Data | Text Data | Image Data | Dempster-Shafer Fusion |
|---|---|---|---|
| Accuracy | 0.93 | 0.86 | 0.97 |

## VI. Conclusion

Dempster-Shafer fusion in social media is a promising technique to combines the features of multiple modalities. This paper presents the data fusion approach in multimedia data for event detection in twitter by using Dempster-Shafer evidence theory. The combined feature vector is experimentally tested and compared with text feature vector only and image feature only, and the result for event detection with combined vector shows better accuracy. Future work will focus on using late fusion for text and image features, and compare it with our method.

## Acknowledgment

The primary author and related research is sponsored by Najran University in Saudi Arabia.

## REFERENCES

[1] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," in IEEE International Conference on Multimedia and Expo (ICME), 2013, pp. 1-6.
[2] S. Zhou, H. Leung, and F. Yao, "Multimedia Data Fusion," Mathematical Problems in Engineering, vol. 2013, 2013.
[3] L. A. Klein, Sensor and data fusion: a tool for information assessment and decision making vol. 324: SPIE press Bellingham^ eWA WA, 2004.
[4] Mark Maybury and S. Walter, "Multimedia Information Extraction," 2008.
[5] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," Multimedia systems, vol. 16, pp. 345-379, 2010.
[6] P. Sharma and M. Kaur, "Multimodal classification using feature level fusion and SVM," Int. J. Comput. Appl, vol. 76, pp. 26-32, 2013.
[7] J.-P. Isson and J. Harriott, Win with advanced business analytics: creating business value from your data: John Wiley & Sons, 2012.
[8] N. Mishra and D. S. Silakari, "Image mining in the context of content based image retrieval: a perspective," IJCSI International Journal of Computer Science Issues, vol. 9, pp. 98-107, 2012.
[9] Y. Kompatsiaris and P. Hobson, Semantic Multimedia and Ontologies: Springer, 2008.
[10] A. Ma, A. Flenner, D. Needell, and A. G. Percus, "Improving Image Clustering using Sparse Text and the Wisdom of the Crowds," arXiv preprint arXiv:1405.2102, 2014.
[11] S. M. Alqhtani, S. Luo, and B. Regan, "Fusing Text and Image for Event Detection in Twitter," arXiv preprint arXiv:1503.03920, 2015.
[12] J. Mao, Multimodal Data Fusion As a Predictior of Missing Information in Social Networks: Arizona State University, 2012.
[13] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," Computational Intelligence, vol. Volume 31, 2013.
[14] G. Shafer, "A Mathematical Theory of Evidence," AMC, vol. 10, p. 12, 1976.
[15] J. Li, S. Luo, and J. S. Jin, "Sensor data fusion for accurate cloud presence prediction using Dempster-Shafer evidence theory," Sensors, vol. 10, pp. 9384-9396, 2010.
[16] Z. Zhang, T. Liu, and W. Zhang, "Novel Paradigm for Constructing Masses in Dempster-Shafer Evidence Theory for Wireless Sensor Network's Multisource Data Fusion," Sensors, vol. 14, pp. 7049-7065, 2014.
[17] J. G. Li, C. N. Zheng, H. W. Xuan, and Y. Jiang, "Data Fusion in Environment Monitoring Systems with Extended Dempster-Shafer Theory," Applied Mechanics and Materials, vol. 543, pp. 1074-1077, 2014.
[18] O. Ozdikis, H. Oguztuzun, and P. Karagoz, "Evidential location estimation for events detected in twitter," in Proceedings of the 7th Workshop on Geographic Information Retrieval, 2013, pp. 9-16.
[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, pp. 513-523, 1988.
[20] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2, pp. 37-52, 1987.
[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, pp. 91-110, 2004.
[22] R. M. Gray, "Vector quantization," ASSP Magazine, IEEE, vol. 1, pp. 4-29, 1984.