MCA-NN: Multiple Correspondence Analysis based Neural Network for Disaster Information Detection

Haiman Tian and Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
{*htian005, chens*}*@cs.fiu.edu*

***Abstract*—This paper proposes a semantic content analysis framework for reliable video event detection. In this work, we target to improve the concept detection results by feeding the learnt results from individual shallow learning models into a generic model to dig out of the similarities in deeper layers. Compared to the deep learning models, the shallow learning models are memorizing rather than understanding the features. The proposed framework tackles the issue in shallow learning by integrating the strength of Multiple Correspondence Analysis (MCA) and Multilayer Perceptron (MLP) neural network. The low-level features are taken as the initial inputs for MCA-based models to abstract higher-level feature values. The output values further involve interaction in the neural network for better understanding. It earns the ability to put forward the arguments. The framework provides final decisions of video classifications by analyzing the decisions of every single frame from the network outputs.**

***Keywords*-Multiple Correspondence Analysis (MCA); Neural Network (NN); disaster information system; video concept detection**

I. Introduction

Disasters take place frequently in recent years. They affect the normal activities of communities with serious losses, which includes economic and environmental losses, as well as human lives [1][2][3]. If there exists an effective disaster information management system [4] that can be triggered immediately to provide the current status report of the hazard to the community, it is possible to improve the ability of stopping unnecessary consequential losses in time. In that case, not only the hazard status but also the preparation or recovery processes are critical to the populace and the community [5][6].

In the Internet age, the volume of multimedia data (including video, audio, image, and text) grows exponentially, carrying a variety of valuable information [7][8][9]. Multimedia data can be accessed from different kinds of devices, making it more convenient for people to get a visual understanding of the situations that they care about [10][11]. When a catastrophe happens, the modern mobile devices become essential and really helpful to capture disaster related multimedia data. The time-sensitive information is no longer restricted to be published to the public by the Emergency Operations Center (EOC) but can be provided through any trustable organizations.

Video semantic concept detection, which aims to explore the rich information in the videos, uses various machine learning and data mining approaches to address this challenge [12][13][14]. In addition, many existing approaches are making every effort to better fill in the gap between the low-level visual features and the high-level concepts [15][16][17].

Unlike the simple disaster concept detection or disaster classification tasks, which attempt to classify the disaster scenes from non-disaster scenes, the information concentrated to one disaster includes the disaster event, damage situation, disaster recovery, disaster effect, and in advance prevention, to name a few [6]. The difficulty increases since all those concepts are surrounding one major premise, which will immensely increase the similarity between the concepts. In the literature, various classifiers have be used to identify the inherent concepts in videos [18], including Multiple Correspondence Analysis (MCA), decision trees [19], etc. However, there is still a large space of improvements. Beyond the shallow learning method, neural networks, like Multilayer Perceptrons (MLP) [20], are considered to target complex learning purposes that achieve the ability to explore in a greater detail.

In this paper, a novel framework of Multiple Correspondence Analysis based Neural Network (MCA-NN) is proposed to address the challenges in shallow learning. It integrates the Feature Affinity based Multiple Correspondence Analysis (FA-MCA) models into one large neural network model. The major contributions of this work are as follows: First, this is the first time that the MCA-based model is applied to separated groups of features and generates higher-level features as the inputs of the deep learning component; Second, the proposed semantic concept detection framework is utilized to decide the video concept instead of frame-based classification; Furthermore, the process of deciding the neural network module is automatic. The most important parameters building the network are obtained from the outputs of the FA-MCA models and the corresponding statistical information.

The rest of this paper is organized as follows. Section II discusses the related work in multimedia data analysis. Section III details the proposed framework with the discussions of each component. In Section IV, the experimental results and performance evaluation are presented. Finally, the last section concludes this paper.

II. Related Work

The traditional multimedia data analysis uses hand-crafted features (low-level features) with simple trainable classifiers which are widely used in various domains [21][22]. Those diversified representatives are converged in a single form and stored for future analysis. High-level features or concepts can be learnt from the raw data using trainable feature extractors. In order to convey a group of low-level features to a proper high-level semantic concept, several approaches can be involved in the procedure, like feature selection [23][24][25], feature extraction [26], and classifier selection [27][28][29].

Feature selection reduces the feature dimension that efficiently speeds up the learning process. Furthermore, the advances in technology have also made it possible to record the multimedia data in higher resolutions. As a double-edged sword, it improves the analysis results distinctively by increasing the feature quality but also slows the analysis process due to the increase in feature quantity.

The learning process is considered as deep learning if it has more than one stage of non-linear feature transformation [30]. Along with transforming the low-level features into mid-level and high-level features, the level of abstraction increases with the hierarchical representations. The MLPs are used as the base of the deep learning architectures, which provide a complex function to determine the feature values in the feedforward direction.

In the MCA-NN framework, the input representations of a low-level feature are transformed into a higher-level value using FA-MCA model training. However, it is one stage feature transformation, which is considered as shallow learning while high-level features are more global and more invariant. To address this issue, it is worth considering the MLP neural network, which takes the transforming features to the predictor.

III. The Proposed Framework

The overall framework is illustrated in Figure 1. It includes three major steps: pre-processing (the upper right panel), training phase (the upper left panel), testing phase (the lower right panel). The output classification results from the network are frame based. The final classification of the framework concludes the single frame decisions for each video to produce the entire video classification.

The pre-processing phase includes key frame extraction and feature extraction, which make the data cleaned and structured. In the training phase, the model is trained using the FA-MCA algorithm for each feature group independently. The low-level features were learnt through each FA-MCA model and transformed into a higher-level feature. Each model produces one ranking score for each instance, and the ranking score is normalized as a new feature that includes a higher level semantic. Followed by the FA-MCA model training, an MLP network is created using the FA-MCA outputs to deeply learn the relationships between high-level features.

The low-level feature value affinities are calculated and accumulated as weighting factors, which will be used in the testing phase to generate the high-level feature value of the testing instances. The low-level feature sets are distributed into different groups for high-level feature value extraction based on the different representation levels (e.g., color space, object space, etc.). For example, from the color space to the object space, the feature groups form a flat structure, indicating that each group is self-structured and relatively independent. Afterward, the outputs from the FA-MCA models are utilized as inputs of hierarchical feature learning network, which makes use of the relationships between independent high-level feature values.

A. Pre-processing

In video analysis, the pre-processing phase for each video is independent while several low-level features are extracted from every frame. To reduce the number of frames in the process, one raw video is separated into different video shots [31]. Only one key frame is selected to represent the video shot, and all the selected key frames are used to cover the whole idea of the video. This process reduces the computation time significantly.

In this paper, several different types of low-level visual features are extracted from the raw data includes Histogram of Oriented Gradient (HOG) [32], Color and Edge Directivity Descriptor (CEDD) [33], Haar-like feature [34], and color space information [35]. Specifically, HOG feature is used for the purpose of object detection, which is computed on a dense grid of uniformly spaced cells and uses overlapping normalization for accuracy improvement. CEDD feature, as it is named, obtains color information and texture information. Haar-like feature is always used in object recognition with Haar wavelets, especially useful in face detection. Color space representations are considered using Hue, Saturation, and Value (HSV), with YCbCr as the supplemental information. As a result, one video is represented by several key frames, and each key frame is composed by several feature values. Hence, the dataset consists of data instances at the frame level with the binary class information. The finalized dataset is then split into training and testing sets using three-fold cross-validation [36] based on the count of videos. In other words, the entire group of key frame instances that belong to one video is assigned to either the training dataset or testing dataset

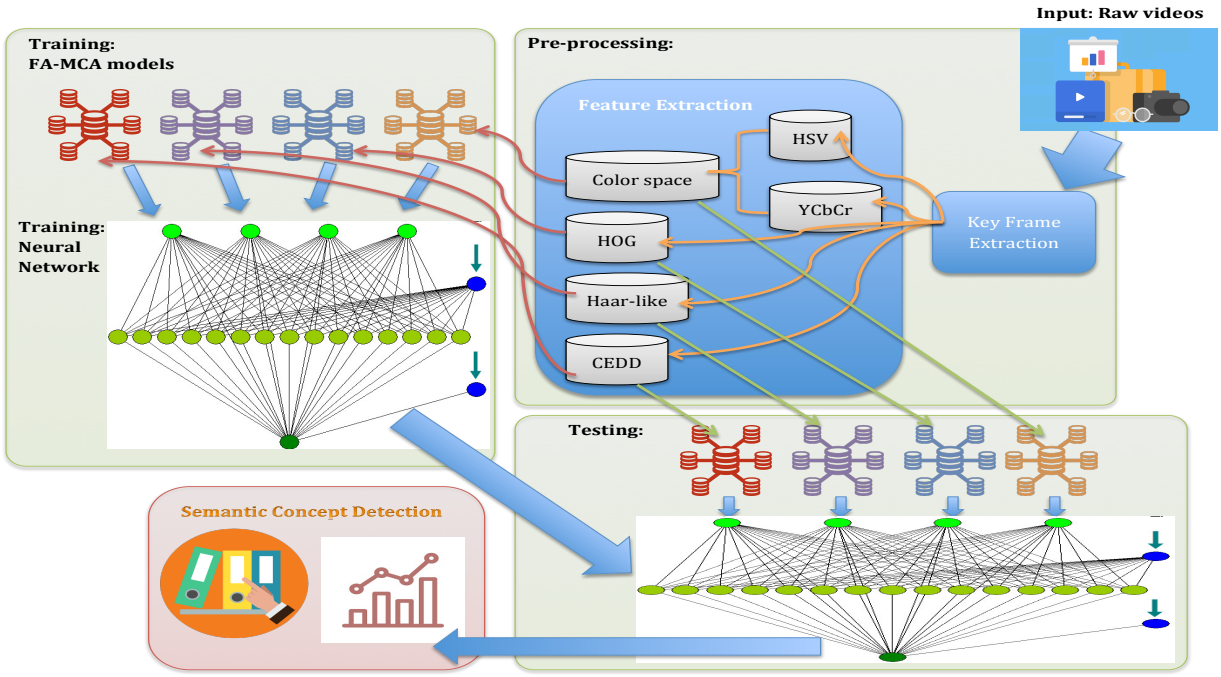


Figure 1: Illustration of the MCA-NN framework

during the separation, in order to preserve the information between frames that represents in each video.

B. Training Phase

In the training phase, there are two key components: low-level feature transformation using FA-MCA and the MLP neural network that takes the transformed feature values as inputs; and feature transformation component uses FA-MCA to produce a single value for each feature group, which leverage the low-level feature group into a more abstract representative value. The calculation bases on a weighting matrix that takes each low-level feature into consideration during learning process. The neural network builds up based on the outputs and the statistics of the training results from each FA-MCA model. The output values are ranking scores for training instances that can be used for classification purpose. However, in this proposed work, the ranking scores are used as higher-level features for the following deep learning process.

To fully utilize the value in each output, the number of hidden layers is decided by the number of input layers. Considering the permutation of N optional input layers, the full permutation of the selection is $N!$. One bias weight (w_{i0}) is included in the total number of hidden neurons and will not be updated during the back propagation, as well as the one counts for the input neurons. For example, in the proposed framework, there are four input layers for different low-level features. As taking variable N as 4, there are 5 input weights in total, which are 4 weights for different inputs plus one bias weight. For the hidden layers, there are $4! = 24$ weights plus one bias weight.

The $tanh$ activation function is used to enable a wider range of output instead of linear activation, the input neurons (a_i) for next layers are calculated base on the following formula:

$$a_i = tanh(net_i) = \frac{e^{net_i} - e^{-net_i}}{e^{net_i} + e^{-net_i}} \quad (1)$$

The $tanh$ function restricts the output between -1 and 1, which can be used to predict the event if the value turns out to be positive or negative. Therefore, the transformed features are normalized between -1 and 1 as well. If the transformed feature is closer to -1, it means the FA-MCA model has learnt that the low-level features for the specific instance are more likely to represent the target concept, vice versa.

In formula 1, net_i is the correspondence neuron output of current layer, which accumulates the weighted output from the previous layers (shows in formula 2), $Pred(i)$ is the set of all neurons j for a connection $j \rightarrow i$ exists, called the set of predecessors.

$$net_i = w_{i0} + \sum_{j \in Pred(i)} w_{ij} a_j \quad (2)$$

The initial weights for the calculations of all the hidden layers use the F1 scores from the FA-MCA training results, which are the values between 0 and 1. In that case, the initial weight will be large if the transformation shows

high confidence by a large F1 score. A smaller weight will be assigned if the confidence of specific FA-MCA training model is lower. Therefore, the transformed high-level features might not be able to carry out a well learnt concept comparing with other features. To obtain better initial weights for each input, the best F1 scores for the training dataset using the FA-MCA models are modified to fit in the range of [-0.5,0.5]. In order to get an initial output between [-1,1], each weight is divided by 2 (E.g., 4 input layers with each weight between [-0.25, 0.25]). The bias weight takes the average F1 score for all low-level feature transformation models and modified it also fit in the same range of initial weight.

For the output layer (Dark green neuron in figure 1), all the weights for calculating the hidden neurons are initiated randomly following the requirement of range in [-0.5, 0.5]. Since each training data set is unique for each concept, select the weights randomly will not restrict too much to the output. However, it needs several rounds of back propagation to compute gradients. The repeating process is set to 10,000 times during experiments for regular runs to have error plummets. The error rate is accumulated by p training instances and calculated based on formula 3 once of the training cycle to determine the learning rate of the output layers for the next cycle, which is used in the process of back propagation.

$$E_{total} = \frac{1}{2}\sum_{i=1}^{p}(target_i - output_i)^2 \tag{3}$$

The weights are updated during the back propagation in order to have the actual output ($output_i$) to be closer to the target output ($target_i$). Namely, minimizing the error for each hidden neuron and the whole network. The changes of weights (w_i) in the output layer calculation affect the total error by taking the partial derivation as following:

$$\frac{\partial E_{total}}{\partial w_i} = \frac{\partial E_{total}}{\partial output_i} * \frac{\partial output_i}{\partial net_i} * \frac{\partial net_{ti}}{\partial w_i} \tag{4}$$

The partial derivative of the activation function is 1 minus the square of the current layer output (shows in formula 5).

$$\frac{\partial output_i}{\partial net_i} = 1 - tanh^2(net_i) \tag{5}$$

The backward calculation of the weight changes for hidden layers is similar but slightly different to account the output of each hidden layer neuron contributes to the output neuron. So every hidden layer weight change is the partial derivative of the total hidden layer input with respect to each weight (w_{ji}), where j is the total number of input neurons:

$$\begin{aligned}\frac{\partial E_{total}}{\partial w_{ji}} =& (\sum_{j} \frac{\partial E_{total}}{\partial output_j} * \frac{\partial output_j}{\partial net_j} * \frac{\partial net_j}{\partial output_{ji}}) \\ & * \frac{\partial output_{ji}}{\partial net_{ji}} * \frac{\partial net_{ji}}{\partial w_{ji}}\end{aligned} \tag{6}$$

Both hidden layers' and output layers' weights are updated during the runs to decrease the error by multiplying by a learning rate, the following formula shows the update step, where w_i^+ represents the updated weight:

$$w_i^+ = w_i - \eta * \frac{\partial E_{total}}{\partial w_i} \tag{7}$$

The learning rates η for both updating functions (output layer and hidden layers) are set to 0.7 empirically at the initial step. However, in some of the training process, 0.7 seems too large to tighten up the errors. It takes so many learning cycles but still could not be able to find a proper prediction value with a low error rate. The proposed framework automatically detects the large error rates after the first 1000 runs as tolerance. If the total error remains greater than 0.01, the learning rate of the output layer will be reduced by 10 times (reset to 0.07). Consequently, since the learning rate affects the duration of the learning process, the training cycle extended two times longer than the original one to acquire an output prediction value with an acceptable error rate.

C. Testing phase

The final weighting matrix generated during the training phase of FA-MCA is used in the testing phase in order to get the final ranking scores for the testing instances. Those ranking scores are responsible for representing the high-level concepts. The ranking procedure starts with adding all feature weights for instance t, and calculates its average value [37].

For the purpose of feeding the variables into the well-trained neural network, all the ranking scores of the testing instances are normalized between [-1,1] as the training instances in order to better represent the value that is similar to the output of the $tanh$ function.

Since the best F1 score for the training data can be calculated by attempting to separate the transformed low-level features into the positive class (containing the target concept) or the negative class (not containing the target concept), the F1 score for each FA-MCA model is recorded as the confidential variable that can be utilized for initializing the MLP weights.

The well-trained MLP network is directly used by feeding all the testing instances one by one to generate the prediction values. As all the weights are updated and fixed during the training phase to optimally derive the positive instances from the negative instances, the testing phase is as easy as running the fixed network to compute the output. Same to the ideal distribution in the training phase, a smaller output value in the range of [-1,1] predicts a positive instance, while a larger output value predicts a negative one. The number 0 is selected as the value to do the classification, which means the instance holding a prediction value smaller than zero will be classified as positive.

(a) Flood

(b) Human Relief

(c) Damage

(d) Training Program

(e) Disaster Recovery

(f) Speak

(g) Interview

Figure 2: Different sample concepts in the dataset

D. Semantic concept detection

As mentioned earlier, the final semantic concept predictions are concluded by the count of the videos. In that case, the output from the neural network, which has the classification results for each individual frame, needs to be integrated to get the finalized decision of each video. The framework takes the classification results of the frames for one video to decide the final classification. By counting the total number of frames for one video that are being tested, the portion of the predicted class (negative or positive) affects the final decision. During the experiments, the portion threshold is set to be 0.6, which means if there are more than 60% of the frames being classified as negative, the video will be classified as negative. Otherwise, the video will be predicted as positive. The negative labeled video means that the target semantic concept is not detected from the tested video. On the contrary, if the video is classified as positive, it means that the concept is detected. The experiments of how to decide the threshold is shown in the following section.

No.	Concepts	Positive Instances	Videos
1	Flood	258	21
2	Human Relief	92	4
3	Damage	281	21
4	Training Program	148	7
5	Disaster Recovery	369	16
6	Speak	1230	145
7	Interview	117	23
	Total	2495	237

Table I: Dataset statistics

IV. Experimental Analysis

A. Dataset Description

In this paper, a specific task of detecting disaster-related semantic concepts is selected using a dataset obtained from the Federal Emergency Management Agency (FEMA) website, although the framework can be used as a general framework that works for various multimedia application domain. The semantic concepts obtained from this website are different from the normal disaster event concepts. It is more useful to examine the effectiveness of the proposed MCA-NN framework that improves the capability of detecting the differences between similar concepts.

The dataset includes more than 200 videos, which contain thousands of key frames that are related to seven different concepts. However, there are still a great amount of similarities between the concepts. The statistics information is shown in Table I that depicts the name, the number of positive instances, and the number of videos of each concept. When the similarity between concepts increases, the task of concept detection becomes more challenging. Meanwhile, a well trained neural network for the transformation of features improves the training and testing performance. These are the reasons and motivation of proposing the MCA-NN framework. As mentioned in Section III-A, the dataset is split using three-fold cross-validation based on the number of videos. In other words, the entire data set is divided into 3 different folds with approximately 1/3 of the videos (one fold) for testing and 2/3 of the videos (two folds) for training.

Figure 2 also depicts the samples of each concept in details on which are the key frames extracted from the videos and used during evaluation process. It is easier to differentiate the concept “Flood” in Figure 2a from the concept “Human Relief” in Figure 2b than to distinguish the concept “Speak” in Figure 2f from the concept “Interview” in Figure 2g.

B. Evaluation Results

The performance evaluation takes the precision, recall, and F1-score values as the criteria [38], which consider the number of positive and negative instances in each class. The F1-score measure is considered as the most valuable comparison metric since it is the trade-offs between the precision and recall values. All the classifiers are tuned to achieve their best performance during the experiment.

	Decision Tree			MLP			MCA-NN		
Concepts	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Flood	70.13%	45.23%	29.77%	**70.27%**	51.03%	34.90%	69.74%	**76.19%**	**57.64%**
Human Relief	1.43%	32.33%	2.73%	1.63%	31.30%	3.07%	**33.99%**	**50.00%**	**23.51%**
Damage	**72.10%**	61.40%	49.13%	5.10%	33.33%	8.87%	64.14%	**71.43%**	**51.20%**
Training Program	**68.60%**	40.87%	17.20%	25.67%	38.56%	15.61%	67.51%	**88.89%**	**61.65%**
Disaster Recovery	70.37%	61.67%	46.47%	**70.57%**	65.83%	49.87%	60.53%	**81.11%**	**56.64%**
Speak	82.93	81.37%	77.67%	78.23%	**95.57%**	83.67%	**86.92%**	93.88%	**88.49%**
Interview	68.77%	40.53%	18.90%	35.53%	36.77%	10.30%	**70.09%**	**77.38%**	**59.01%**
AVERAGE	62.05%	51.91%	34.55%	41.00%	50.34%	29.47%	**64.70%**	**76.98%**	**56.88%**

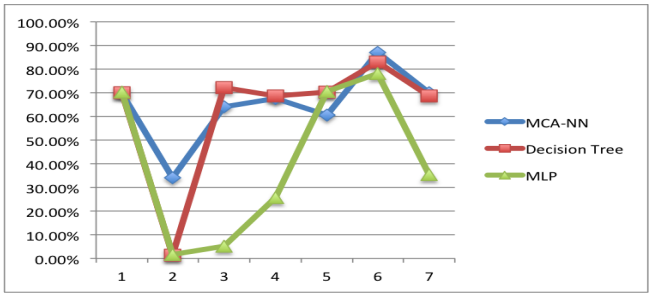
Table II: Performance evaluation results on a disaster dataset

The proposed framework shows the best performance on average in comparison with the decision tree and MLP classifiers (available in WEKA [39]). The performance by each comparison criterion is illustrated in Figure 3. Each plot takes the concept id as the x-axis and the percentage evaluation result as the y-axis. The concept id that refers to a different concept name can be found in Table I. It is clear that, during the comparison of each criterion, the proposed method wins most of them, especially in the comparison of the recall and F1-score values, which are in Figure 3b and Figure 3c, respectively.

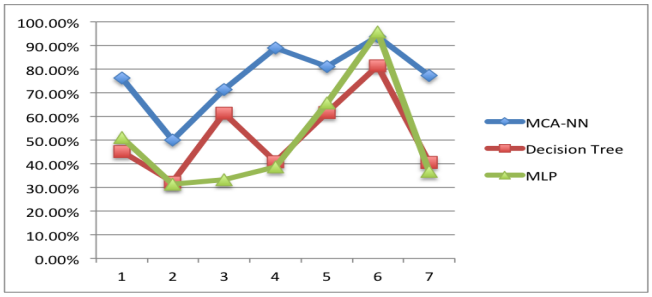
Table II presents the experimental results in details. As can be seen from this table, the improvement of the average F1-score is more than 27% when comparing to MLP. Compared to the Decision Tree, the average results (precision, recall, F1 score) improve 2.65%, 25.07% and 22.33%, respectively. Although the MLP recall performs nearly two percent better than MCA-NN for one of the concepts (i.e., Speak), it does not get the best F1 score, which means it takes as many instances as positive; while more negative instances are wrongly classified. Also, it shows poor performance when the number of positive instances is very small (i.e., imbalanced data). However, MCA-NN performs well, no matter whether the number of positive instances is large or small in a dataset.

Additionally, since we prefer to recognize as many related events as possible for the purpose of disaster information analysis, the recall values earn more attention when comparing to the precision values. However, blindly increasing the number of positive instances in the classification process could only bring a higher recall value. A better F1 score relies on a more accurate classification framework. In other words, the increasing recall values at the cost of the precision values would not be able to get a stable F1 score in the experiments.

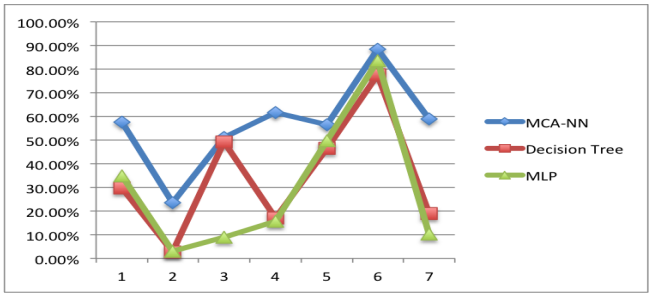
Figure 4 shows the experiments on selecting the best threshold of making decisions for entire video classifications. It is clear that the precisions are affected slightly during the test. The rightmost three bars, which represents taking 0.6 as the threshold that is used for all the experimental results depicted in Table II, show the best recall and F1 score values in this test. From the test, we can conclude that since the precision values would not be greatly affected by the threshold, it would be better to increase the threshold in order to get the best recall and F1-score values. The bar chart shows a gradually increasing trend for both recall and F1-score values, accompanying with an increasing threshold. However, when the threshold comes to 0.7, the precision value suddenly dropped to 0 in the test. So the final threshold is determined to be 0.6.



(a) Precision



(b) Recall



(c) F1

Figure 3: Different evaluation criteria results

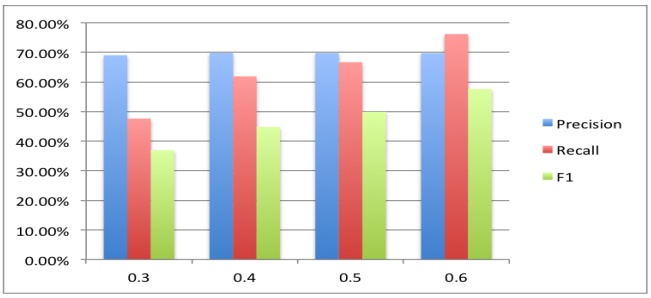


Figure 4: The experimental results for deciding the video classification threshold

V. Conclusion

Disaster-related concept detection includes disaster event detection, disaster preparation training, disaster recovery, and disaster damage situation, to name a few. Since it does not limit to the straight forward disaster events, the concepts that need to be utilized are varied for the aim of managing the disaster information. Since the correlations between those concepts are higher than the diverse disaster events, it makes the classification task more challenging. To tackle this challenge, in this paper, the MCA-NN framework is proposed to convey the low-level features into the higher-level feature values through the FA-MCA models, considering the relationship between the features within each feature group. The shallow network learned and transformed features were used as the input for a deeper learning neural network for further training purpose. As a result, critical low-level features are memorized and depicted as the higher-level features. Consequently, the higher-level features are explored in details to better understand the concepts.

Comparing with the decision tree and MLP classifiers, the experimental results show significant improvements for all the evaluation criteria, which means that the proposed framework successfully transformed the low-level features and truly learnt the concepts when differentiating the inter-related concepts. However, there is still some improvements that can be further carried out.

In the future, this framework will be further extended and tested for more concept detection applications. It is worth considering to do more research on the randomly assigned initial weights in order to reduce the repeating cycles. Other neural networks and back propagation algorithms can be utilized to better fulfill the deep learning purpose.

Acknowledgment

This research is partially supported by DHSs VACCINE Center under Award Number 2009-ST-061-CI0001 and NSF HRD-0833093, HRD-1547798, CNS-1126619, and CNS-1461926. This is contribution number 821 from the Southeast Environmental Research Center in the Institute of Water & Environment at Florida International University.

References

[1] D. Zhang, L. Zhou, and J. F. Nunamaker Jr, "A knowledge management framework for the support of decision making in humanitarian assistance/disaster relief," *Knowledge and Information Systems*, vol. 4, no. 3, pp. 370–385, 2002.

[2] Y. Yang and S.-C. Chen, "Disaster image filtering and summarization based on multi-layered affinity propagation," in *2012 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2012, pp. 100–103.

[3] V. Hristidis, S.-C. Chen, T. Li, S. Luis, and Y. Deng, "Survey of data management and analysis in disaster situations," *Journal of Systems and Software*, vol. 83, no. 10, pp. 1701–1714, 2010.

[4] Y. Yang, W. Lu, J. Domack, T. Li, S.-C. Chen, S. Luis, and J. K. Navlakha, "Madis: A multimedia-aided disaster information integration system for emergency management," in *2012 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*. IEEE, 2012, pp. 233–241.

[5] S. Luis, F. C. Fleites, Y. Yang, H.-Y. Ha, and S.-C. Chen, "A visual analytics multimedia mobile system for emergency response," in *2011 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2011, pp. 337–338.

[6] Y. Yang, H.-Y. Ha, F. Fleites, S.-C. Chen, and S. Luis, "Hierarchical disaster image classification for situation report enhancement," in *2011 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2011, pp. 181–186.

[7] S.-C. Chen, R. L. Kashyap, and A. Ghafoor, *Semantic models for multimedia database searching and browsing*. Springer Science & Business Media, 2000, vol. 21.

[8] S.-C. Chen and R. Kashyap, "Temporal and spatial semantic models for multimedia presentations," in *1997 International Symposium on Multimedia Information Processing*, 1997, pp. 441–446.

[9] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen, "Category cluster discovery from distributed www directories," *Information Sciences*, vol. 155, no. 3, pp. 181–197, 2003.

[10] Y. Yang, H.-Y. Ha, F. C. Fleites, and S.-C. Chen, "A multimedia semantic retrieval mobile system based on hcfgs," *IEEE MultiMedia*, vol. 21, no. 1, pp. 36–46, 2014.

[11] H. Tian and S.-C. Chen, "A video-aided semantic analytics system for disaster information integration," in *IEEE International Conference on Multimedia Big Data (BigMM)*, 2017.

[12] H.-Y. Ha, Y. Yang, S. Pouyanfar, H. Tian, and S.-C. Chen, "Correlation-based deep learning for multimedia semantic concept detection," in *International Conference on Web Information Systems Engineering*. Springer, 2015, pp. 473–487.

[13] H.-Y. Ha, F. C. Fleites, S.-C. Chen, and M. Chen, "Correlation-based re-ranking for semantic concept detection," in *2014 IEEE 15th International Conference on Information Reuse and Integration (IRI)*. IEEE, 2014, pp. 765–770.

[14] L. Lin, M.-L. Shyu, and S.-C. Chen, "Rule-based semantic concept classification from large-scale video collections," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 4, no. 1, pp. 46–67, 2013.

[15] S.-C. Chen, "Multimedia databases and data management: a survey," *International Journal of Multimedia Data Engineering and Management*, vol. 1, no. 1, pp. 1–11, January-March 2010.

[16] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn, "Image database retrieval utilizing affinity relationships," in *Proceedings of the 1st ACM international workshop on Multimedia databases*. ACM, 2003, pp. 78–85.

[17] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Video semantic concept discovery using multimodal-based association classification," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 859–862.

[18] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen, "Weighted subspace filtering and ranking algorithms for video concept retrieval," *IEEE MultiMedia*, vol. 18, no. 3, pp. 32–43, 2011.

[19] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems Man and Cybernetics*, pp. 660–674, 1990.

[20] P. Gallinari, S. Thiria, F. Badran, and F. Fogelman-Soulie, "On the relations between discriminant analysis and multilayer perceptrons," *neural networks*, vol. 4, no. 3, pp. 349–360, 1991.

[21] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang, "User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval." in *MDM/KDD*. Citeseer, 2002, pp. 100–108.

[22] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying overlapped objects for video indexing and modeling in multimedia database systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 04, pp. 715–734, 2001.

[23] Q. Zhu, L. Lin, M.-L. Shyu, and S.-C. Chen, "Feature selection using correlation and reliability based scoring metric for video semantic detection," in *2010 IEEE Fourth International Conference on Semantic Computing (ICSC)*. IEEE, 2010, pp. 462–469.

[24] J. Fan, H. Luo, J. Xiao, and L. Wu, "Semantic video classification and feature subset selection under context and concept uncertainty," in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2004, pp. 192–201.

[25] X.-w. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 124–132.

[26] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2018–2025.

[27] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 188–197.

[28] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, "Spatio-temporal analysis for human action detection and recognition in uncontrolled environments," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 6, no. 1, pp. 1–18, 2015.

[29] Q. Zhu, M.-L. Shyu, and S.-C. Chen, "Discriminative learning-assisted video semantic concept classification," *Multimedia Security: Watermarking, Steganography, and Forensics*, p. 31, 2012.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, no. 2, pp. 122–128, 1996.

[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[33] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *International Conference on Computer Vision Systems*. Springer, 2008, pp. 312–322.

[34] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *2002 International Conference on Image Processing.*, vol. 1. IEEE, 2002, pp. I–900.

[35] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," in *Proceedings of the 2002 International Conference on Image Processing.*, vol. 2. IEEE, 2002, pp. II–589.

[36] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Artificial Intelligence Journal (IJCAI)*, vol. 2, 1995, pp. 1137–1143.

[37] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Correlation-based video semantic concept detection using multiple correspondence analysis," in *Tenth IEEE International Symposium on Multimedia (ISM)*. IEEE, 2008, pp. 316–321.

[38] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *International Journal of Machine Learning Technology*, 2011.

[39] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: A machine learning workbench," in *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, 1994.* IEEE, 1994, pp. 357–361.