Semantic Event Detection Using Ensemble Deep Learning

Samira Pouyanfar and Shu-Ching Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
{*spouy001, chens*}*@cs.fiu.edu*

***Abstract*—Numerous deep learning architectures have been designed for a variety of tasks in the past few years. However, it is almost impossible for one model to work well for all kinds of scenarios and datasets. Therefore, we present an ensemble deep learning framework in this paper, which not only decreases the information loss and over-fitting problems caused by single models, but also overcomes the imbalanced data issue in multimedia big data. First, a suite of deep learning algorithms are utilized for deep feature selection. Thereafter, an enhanced ensemble algorithm is developed based on the performance of each single Support Vector Machine classifier on each deep feature set. We evaluate our proposed ensemble deep learning framework on a large and highly imbalanced video dataset containing natural disaster events. Experimental results demonstrate the effectiveness of the proposed framework for semantic event detection, and show how it outperforms several state-of-the-art deep learning architectures, as well as handcrafted features integrated with ensemble and non-ensemble algorithms.**

***Keywords*-Deep learning; Ensemble learning; Imbalanced data; Semantic event detection; Multimedia big data.**

I. Introduction

Over the last decade, social networks and multimedia sources such as Twitter, YouTube, and Facebook have generated a significant amount of digital data. For example, over hundred hours of videos are uploaded to YouTube every minute in a day. Due to such multimedia data explosion, as well as its rich and significant contents, it is considered as a valuable source of data in many research studies [1], [2]. Video semantic event detection is one of the main applications of multimedia management systems. Recently, many researchers have tried to detect the most interesting events and concepts from videos [3], [4]. Criminal event detection from video and audio data, natural disaster retrieval from video data, and interesting event detection in a sport game are a few examples of video semantic event detection.

However, there are several challenges needed to be addressed in multimedia semantic analysis, including how to analyze such huge volume and variety of data in an efficient manner, and how to handle data with a non-uniform distribution. The latter is known as imbalanced data problem, which has been commonly seen in video event detection scenarios. For example, suppose one is looking for video shots containing natural disaster information among thousands of videos in YouTube where meta-data and textual information may not be reliable and accurate. This example shows that the skewed distribution of the major class (non-disaster video shots) and minor or interesting class (videos containing disaster information). This rareness of interesting events in videos makes the detection task more challenging. Currently, the class imbalance issue has been studied by many researchers in the literature [5], [6], [7]. Nevertheless, conventional learning approaches are still biased toward the majority classes.

In the past few years, deep learning has attracted lots of attention in both academia and industry [8], [9], [10]. It is one of the significant breakthrough techniques in data mining and machine learning algorithms [11]. Using a cascade of layers in a deep graph architecture, composed of multifold linear and non-linear transformations, deep learning intents to model very high-level data abstractions. The recent explosion of deep learning studies has led to significant advances and improvements in multimedia management systems. Although deep learning techniques have been applied to lots of research studies in recent years, there is still limited work focusing on the imbalanced data problem in multimedia data.

Multi-classifier fusion is another hot topic in data mining and machine learning because one single classifier can be hardly applied to all scenarios and usually cannot handle imbalanced and big multimedia data due to over-fitting, information loss, and additional bias [12]. Inspired by the fast progress and achievements of deep learning, this paper leverages deep learning techniques for video feature analysis with the application to semantic event detection. In addition, due to the great success of ensemble learning techniques in machine learning, an enhanced ensemble deep learning is proposed in this paper to improve the event detection in imbalanced multimedia data.

The remainder of the paper is organized as follows. In section II, an overview of the state-of-the-art research in imbalanced multimedia analysis is provided. Section III discusses the details of the proposed ensemble deep learning framework. In section IV, a comprehensive experimental analysis is presented. Lastly, we conclude the paper in section V.

II. Related Work

Imbalanced data has been widely seen in many real world applications [5], such as activity recognition, cancer prediction, banking fraud detection, and video mining [1], [13], to name a few. The imbalanced data solutions can be grouped into three categories [14]: (1) Sampling methods which modify data distributions in a way to generate more balanced data, (2) Kernel-based and active learning methods which utilize robust classifications techniques to naturally handle imbalanced learning, and (3) Cost-sensitive methods which apply a penalty for misclassification of instances from one class to another. Among them, the integration of ensemble learning techniques with imbalanced data solutions have shown significant successes in the past years [12], [15].

Another hot topic in multimedia data is how to employ several feature extraction techniques to improve the final detection results. Ha et al. [16] presented a multi-modality fusion technique for multimedia semantic retrieval. Specifically, the correlation between feature pairs is calculated to reduce the feature space by eliminating features with low correlation toward others. Thereafter, features are grouped using Hidden Coherent Feature Groups (HCFGs) technique [17]. Finally, multiple classifiers are trained for all feature groups and the scores generated by each classifier are fused for final event retrieval. In another work, Liu et al. [18] proposed a new feature representation method by integrating spatial and temporal information from video sequences. Using the optical flow field and Harris3D corner detector, as well as a boosting ensemble algorithm based on two classifier models (sparse representation and hamming distance classifiers), the authors successfully improved the performance of human action recognition.

Deep learning is not a new topic and has a long history in artificial intelligence [11]. Convolutional Neural Networks (CNNs) [19], for instance, have improved traditional feed-forward neural networks in 1990s, especially in image processing, by constraining the complexity of networks using local weight sharing topology. Traditional neural network techniques are difficult to interpret due to their black-box nature and they are also very prone to over-fitting [9]. In contrast, new deep learning algorithms are more interpretable because of their strong local modeling. In addition, as new ideas, algorithms, and network architectures have been designed in the last few years, deep learning has shown significant advances mainly in image recognition and object detection.

As a single classifier may not be able to handle large datasets with multiple feature sources, ensemble algorithms have attracted lots of attention in the literature, which can be utilized to enhance the classification performance by taking advantages of multiple classifiers. A positive enhanced ensemble algorithm which handles imbalanced data in video event retrieval is presented [12]. Their proposed framework

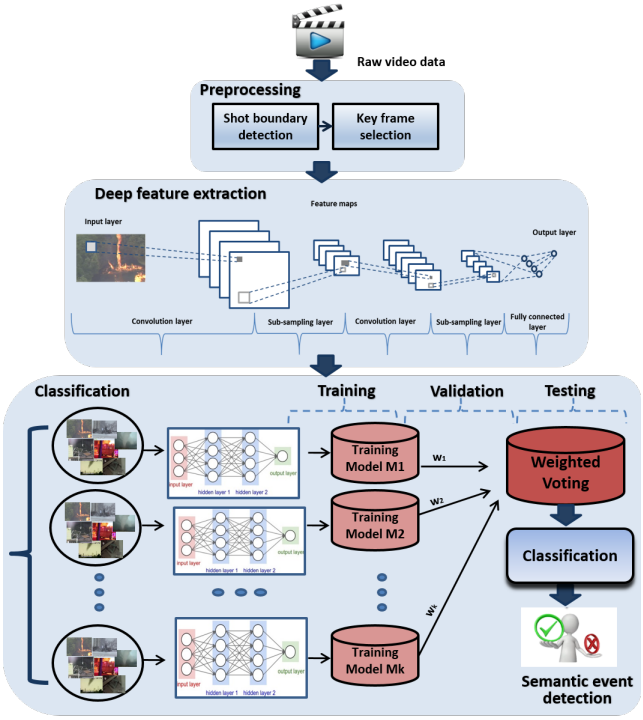

Figure 1: The proposed ensemble deep learning framework

combines a sampling-based method with a classifier fusion algorithm to enhance the detection of interesting events (minor classes) in an imbalanced sport video dataset. An ensemble neural network is proposed in [20]. Using a bootstrapped sampling approach along with a group of neural networks, the rare event issue is alleviated. The framework was also evaluated using a large set of soccer videos with the purpose of corner event detection.

III. Ensemble Deep Learning Framework

In this paper, we propose a mixture of deep learning feature extractors integrated with an enhanced ensemble algorithm. The framework is shown in Figure 1, which not only improves the performance of event detection from videos, but also avoids over-fitting and information losses. The proposed framework is divided into three main modules: (1) preprocessing, (2) deep feature extraction, and (3) classification including training, validation, and testing.

A. Preprocessing

As the preprocessing module is domain specific, different routines may be needed for different applications, such as audio, image, video, and textual analysis. In this study, we apply an automatic shot boundary detection approach [21] on the raw video. This unsupervised algorithm is mainly based on the object tracking and image segmentation techniques.

After all shots are obtained from the raw videos, the first frame of each shot is selected as a keyframe because it is the most distinctive one.

B. Deep Feature Extraction

Deep learning is an emerging research topic which has been advanced tremendously during the last five years. One of the main applications of deep learning is how to generate useful and discriminative features from raw data. In the last decade, researchers have developed various handcrafted features for visual recognition tasks [22]. HOG [23], CEDD [24], and SIFT [25] are few examples of powerful features that have been widely used in computer vision. However, the progress of handcrafted features has slowed down during 2010-2012 and new deep learning architectures have exceedingly raised the performance levels [26]. Therefore, in this paper, we apply various rich and deep feature extraction models based on the CNN algorithm.

CNNs [19] are variations of MultiLayer Perceptron (MLP) networks with the difference in their local connections. The main idea is to have a locally connected network, inspired by the localized biological neurons in animals visual cortex. It contains a complex set of cells which locally filters input data to extract the rich and deep spatially-local correlations in images. A convolutional network generally includes three main layers: (a) stacked convolutional layers, (b) sub-sampling or pooling layers, and (c) fully connected layers [27] as shown in the deep feature extraction module in Figure 1. In the convolutional layer, a number of feature maps are generated by iteratively applying a function across local-region of the whole input. In other words, the input data is convoluted with linear filters followed by nonlinear activation functions. The k^{th} feature map at a given layer is denoted as x_{ij}^k (given in Equation 1), where i and j are the input dimensions, x_{ij}^{k-1} is the input data from the previous layer, f is an activation function (e.g., sigmoid, tanh, etc.), and filters of the k^{th} layer are determined by W_{ij}^k (weights) and b_j^k (bias). A pooling layer is a nonlinear down-sampling, and is located after each convolutional layer. It reduces the number of feature maps by introducing spareness and provides additional robustness to the network. This layer takes a small block from the previous convolutional layer and produces a single output as shown in Equation 2, where $down(.)$ is a subsampling function (e.g., max, average, etc.) and β_{ij}^k is a multiplicative bias.

$$x_{ij}^k = f((W_{ij}^k * x_{ij}^{k-1}) + b_j^k); \quad (1)$$

$$x_{ij}^k = f(\beta_{ij}^k down(x_{ij}^{k-1}) + b_j^k). \quad (2)$$

The last layer of CNNs is called fully-connected layer which is responsible for the high-level reasoning in the network. Similar to regular neural networks, in this layer, all activations in the previous layer are fully connected to a single neuron. The set of all feature maps at the last convolutional-subsampling layers are the input to the first fully connected layer.

In this paper, we utilize four advanced and successful deep learning architectures for visual feature extraction as explained below.

- AlexNet [8]: the first work that made CNNs popular in image processing. It significantly outperforms the second runner-up (over 10%) in ILSVRC 2012. The Alexnet architecture is very similar to CNNs, but with larger, deeper, and stacked convolutional layers followed by pooling layers.
- CaffeNet [28]: a replication of AlexNet with some improvements, and developed and trained by the Berkeley Vision and Learning Center (BVLC). It is not trained with the relighting data-augmentation and the pooling layer is done before normalization. This reference model is trained on Image-Net dataset as explained in [29].
- R-CNN [26]: mainly used for object detection tasks. It improved the performance results by over 30% compared to the best results on PASCAL VOC 2012. First, it generates candidate regions by leveraging bounding box segmentation with low-level features, and then applies CNN classifiers to detect objects at those specific locations.
- GooglNet [10]: a deeper and wider network rather than AlexNet, and developed by a Google team. It contains 22 layers of a deep network which utilizes the extra sparsity of layers. This framework, also known as "Inception architecture", attempts to find more optimal locality and repeats it spatially. It has shown its promising performance in ILSVRC 2014 on ImageNet dataset by wining the first place in two object detection and classification tasks .

C. Classification

As ensemble methods alleviate the over-fitting problem and increase the performance results, an enhanced ensemble deep learning algorithm is proposed in this paper. After feature extraction, we analyze the extracted deep features and measure the importance of each feature set. In addition, how to optimally integrated the trained models in an effective way is a key issue. For this purpose, we employ the proposed enhanced ensemble method to adjust the weight coefficients for the classification module (shown in Figure 1) which depicts the multi-layer architecture of our learning method. In this module, there are *k* models, each trained on a feature set, and the performance of each model is considered to adjust the weights of the weak classifiers. It contains two main steps: deep ensemble learning and testing.

1) Deep Ensemble Learning: Algorithm 1 illustrates the training procedure of the proposed deep ensemble learning step. First, the dataset is split into three categories: training T, validation V, and testing T'. T is defined as

$T = \{(t_1, c_1), (t_2, c_2), ..., (t_N, c_N)\}$, where t_i is the i^{th} training instance, N is the total number of training instances, and $c_i \in \{0, 1\}$ is the class for a binary classification task. In addition, the feature sets extracted from all deep learning algorithms are stored in Fr which is another input of the training algorithm.

The proposed ensemble learning is basically constructed based on a set of weak learners or models $M = \{M_j,\ j = 1, 2, \cdots, k\}$ as shown in Lines 1-3 of Algorithm 1, where k is the number of total weak learners. In this paper, we utilize linear Support Vector Machine (SVM) as the weak learner which has been widely used for deep learning classification [30]. Each classifier model M_j is trained using the training instances. After that, we evaluate each model using the validation set V as shown in Lines 4-7 of Algorithm 1. For this purpose, we utilize the F1 measure (the weighted average value of precision and recall) which is a number between 0 (the worst case) and 1 (the best case). Afterward, using the $F1_j$ measure for each trained model M_j, the weight of each model is calculated using Equation 3.

$$W_j = \frac{F1_j}{\sum_{j=1}^{k} F1_j}. \tag{3}$$

This probability gives higher weights to the models that are confident about their prediction. Finally, the weight factors W_j and the trained models M_j are outputted for further classification analysis.

Algorithm 1 Training of Ensemble Deep Learning

Input: Training instances $T\{(t_i, c_i),\ i = 1, 2, \cdots, N\}$, Validation instances $V\{(v_i, c_i),\ i = 1, 2, \cdots, N_2\}$, Feature set $Fr = \{F_j,\ j = 1, 2, \cdots, k\}$.
Output: Weight matrix W_j, Trained models M_j.

1: **for all** $F_j \in Fr(j = 1, \cdots, k)$ **do**
2: $\quad M_j \leftarrow \text{SVM}(T, F_j)$;
3: **end for**
4: **for all** $F_j \in Fr(j = 1, \cdots, k)$ **do**
5: $\quad F1_j \leftarrow \text{VALIDATE}(V, F_j)$;
6: $\quad W_j = \frac{F1_j}{\sum_{j=1}^{k} F1_j}$;
7: **end for**
8: **return** W_j, M_j

2) Testing: In the testing step, a weighted sum of the weak learner results from the k trained models is used to predict the label of each testing instance (as illustrated in Algorithm 2). The inputs of this step include testing data T' and the corresponding features Fr, as well as all the trained models M_j and their assigned weights W_j. In Lines 2-4 of Algorithm 2, the labels L_j $(j = 1, \cdots, k)$ generated by the j^{th} weak learner is calculated for each testing instance. Then, the final predicted label PL_i is generated as shown in Line 5 of Algorithm 2.

Algorithm 2 Testing of Ensemble Deep Learning

Input: Testing instances $T'\{(t'_i),\ i = 1, 2, \cdots, N_3\}$, Feature set $Fr = \{F_j,\ j = 1, 2, \cdots, k\}$, Trained models M_j, Weight matrix W_j.
Output: Predicted labels PL_i.

1: **for all** $t'_i \in T'(i = 1, \cdots, N_3)$ **do**
2: $\quad$ **for all** $F_j \in Fr(j = 1, \cdots, k)$ **do**
3: $\quad\quad L_j \leftarrow M_j(t'_i, F_j)$;
4: $\quad$ **end for**
5: $\quad PL_i = \begin{cases} 1 & if \sum_{j=1}^{k} L_j * W_j \geq \frac{1}{2}; \\ 0 & otherwise \end{cases}$
6: **end for**
7: **return** PL_i

IV. Experimental Analysis

A. Experimental Setup

In this paper, we evaluate our proposed framework using the dataset described in [31] which contains about 80 YouTube videos. Seven different natural disaster events, including flood, damage, fire, mud-rock, tornado, lightening, and snow, are selected. Our purpose is to detect these events from a large set of video frames (almost 7000 shots), where the average fraction of the positive to the negative instances (P/N ratio) is 0.051, which shows the imbalanced data distribution in this dataset.

When the data is imbalanced, how to evaluate the framework is very important and critical because the accuracy or other similar criteria that consider the performance of both negative and positive classes are not reliable. Thus, we evaluate our framework using common metrics for imbalanced data - Precision, Recall, and F1 measure.

Caffe [28] is a convolutional framework for the state-of-the-art deep learning approaches. In addition, it includes a set of pre-trained reference models, such as R-CNN, GoogleNet, and AlexNet, to name a few. In this paper, we extract several sets of semantic features from images using the well-known Caffe reference models. More Specifically, four pre-trained deep learning reference models are utilized for feature analysis as explained in section III-B. We extract features from the last fully-connected layer of each model. For example, layer "fc8" of CaffeNet and AlexNet, "loss3" of GoogleNet, and "fc-rcnn" of R-CNN are used as the output layer of our feature extractors. R-CNN generates 200 feature element vectors, while other three reference models generate 1000-dimension feature vectors each.

B. Experimental Results

Our proposed Ensemble Deep Learning (EDL) framework is compared with two sets of algorithms. The first group uses the handcrafted features, such as HOG, CEDD, color histogram, texture, and wavelet. In total, 707 visual features are extracted from each keyframe. The second group uses

features generated by deep learning. We apply several classifiers such as Decision Tree (DT), Multiple Correspondence Analysis (MCA) [31], and an ensemble algorithm called Boosting for handcrafted features. We also use the SVM classifier for the second group as it has shown a promising performance when it is integrated with deep learning techniques. All the classifiers and deep learning approaches are tuned to reach to their best results on our dataset and they are evaluated through the 3-fold cross validation.

The average performance (precision, recall, and F1-score) of various feature sets integrated with different classifiers are shown in Table I. As can be inferred from the table, the proposed EDL not only improves the classification performance compared to all the well-known deep learning algorithms, but also beats all exiting classifiers that utilized engineering features. In the first group (handcrafted features), the ensemble algorithm (boosting) and SVM show the highest results in terms of F1-score. While, in the second group (deep learning features), AlexNet has the highest F1 score. Although SVM has the highest precision compared to the other algorithms, the low recall value decreases its overall performance. Therefore, we utilize this classification as well as the proposed ensemble method to improve the overall performance using deep feature sets.

A visualized performance comparison is also shown in Figure 2. In this plot, the F1 score of each deep learning algorithm on each disaster event is depicted. As can be seen from the figure, the proposed EDL framework outperforms all the state-of-the-art deep learning techniques for all disaster events. R-CNN has the lowest performance for almost all events, which can be due to its architecture which is designed for region-based object detection, not frame-based semantic event detection. CaffeNet and GoogleNet have very close average performances. CaffeNet has much higher F1 scores for damage and tornado, while GoogleNet significantly outperforms CaffeNet in fire and snow events. Among the four algorithms, AlexNet has almost the best results on all events except lightening and snow. Since our proposed method leverages the four algorithms in an intelligent manner, it successfully improves the performance for all semantic events.

In summary, the experimental results show the high superiority and effectiveness of our proposed framework, compared to various novel data mining algorithms.

V. Conclusion

In the paper, a novel ensemble deep classifier is proposed which fuses the results from several weak learners and different deep feature sets. The proposed framework is designed to handle the imbalanced data problem in multimedia systems, which is very common and unavoidable in current real world applications. Specifically, it is applied to the detection of semantic events from videos. Several experiments have been conducted to evaluate the performance of the proposed framework by comparing to several state-of-the-art deep learning and existing machine learning algorithms. The experimental results demonstrate the effectiveness of the proposed framework for video event detection.

Table I: Average performance of various feature sets and classifiers on the disaster dataset

Features	Classifier	precision	recall	F1-score
handcrafted	DT	0.816	0.823	0.819
handcrafted	MCA	0.894	0.720	0.782
handcrafted	Boosting	0.910	0.841	0.867
handcrafted	SVM	0.957	0.802	0.868
R-CNN	SVM	0.930	0.722	0.794
GoogleNet	SVM	0.918	0.840	0.875
CaffeNet	SVM	0.919	0.840	0.876
AlexNet	SVM	0.924	0.859	0.888
deep features	EDL	0.949	0.883	**0.913**

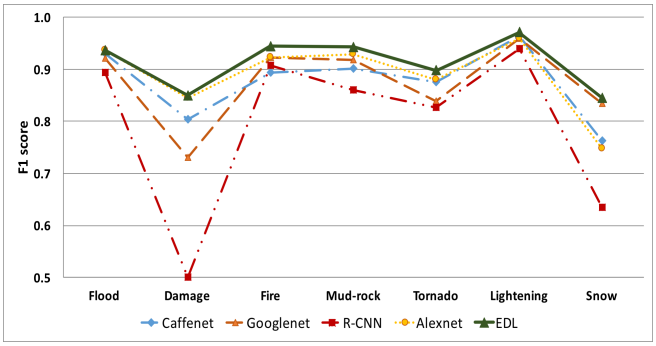

Figure 2: Performance evaluation for different concepts on the disaster dataset

Acknowledgment

This research is partially supported by DHS's VACCINE Center under Award Number 2009-ST-061-CI0001 and NSF HRD-0833093, HRD-1547798, CNS-1126619, and CNS-1461926.

References

[1] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen, "Video semantic event/concept detection using a subspace-based multimedia data mining framework," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 252–259, 2008.

[2] S.-C. Chen, M.-L. Shyu, and C. Zhang, "An intelligent framework for spatio-temporal vehicle tracking," in *Proceedings of the 4th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 2001, pp. 213–218.

[3] 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.

[4] X. Chen, C. Zhang, S.-C. Chen, and S. Rubin, "A human-centered multiple instance learning framework for semantic video retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 2, pp. 228–233, 2009.

[5] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, pp. 1–12, 2016.

[6] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen, "Effective feature space reduction with imbalanced data for semantic concept detection," in *IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC)*. IEEE, 2008, pp. 262–269.

[7] 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.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[11] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 157–166.

[12] Y. Yang and S.-C. Chen, "Ensemble learning from imbalanced data set for video event detection," in *IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2015, pp. 82–89.

[13] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A latent semantic indexing based method for solving multiple instance learning problem in region-based image retrieval," in *Seventh IEEE International Symposium on Multimedia (ISM'05)*. IEEE, 2005, pp. 8–pp.

[14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[15] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2009.

[16] H.-Y. Ha, Y. Yang, F. C. Fleites, and S.-C. Chen, "Correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.

[17] Y. Yang, "Exploring hidden coherent feature groups and temporal semantics for multimedia big data analysis," Ph.D. dissertation, Florida International University, 2015.

[18] 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.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] M. Chen, C. Zhang, and S.-C. Chen, "Semantic event extraction using neural network ensembles," in *International Conference on Semantic Computing (ICSC 2007)*. IEEE, 2007, pp. 575–580.

[21] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative shot boundary detection for video indexing," *Video data management and information retrieval*, pp. 217–236, 2005.

[22] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht, "Image retrieval by color, texture, and spatial information," *Proceedings of the 8th International Conference on Distributed Multimedia Systems (DMS'2002)*, pp. 152–159, 2002.

[23] 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.

[24] 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.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[27] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification," in *IEEE International Symposium on Multimedia (ISM)*, 2015, pp. 483–488.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[29] "Brewing imagenet," retrieved at: 2016-08-09. [Online]. Available: http://caffe.berkeleyvision.org/gathered/examples/imagenet.html

[30] Z. Ge, C. McCool, and P. Corke, "Content specific feature learning for fine-grained plant classification," in *Working notes of CLEF 2015 conference*, 2015.

[31] S. Pouyanfar and S.-C. Chen, "Semantic concept detection using weighted discretization multiple correspondence analysis for disaster information management," in *The 17th IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2016, pp. 556–564.