International Journal of Semantic Computing

Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning

SAMIRA POUYANFAR and SHU-CHING CHEN
School of Computing and Information Sciences
Florida International University
Miami, Florida 33199, USA
spouy001@cs.fiu.edu, chens@cs.fiu.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

With the explosion of multimedia data, semantic event detection from videos has become a demanding and challenging topic. In addition, when the data has a skewed data distribution, interesting event detection also needs to address the data imbalance problem. The recent proliferation of deep learning has made it an essential part of many Artificial Intelligence (AI) systems. Till now, various deep learning architectures have been proposed for numerous applications such as Natural Language Processing (NLP) and image processing. Nonetheless, it is still impracticable for a single model to work well for different applications. Hence, in this paper, a new ensemble deep learning framework is proposed which can be utilized in various scenarios and datasets. The proposed framework is able to handle the over-fitting issue as well as the information losses caused by single models. Moreover, it alleviates the imbalanced data problem in real-world multimedia data. The whole framework includes a suite of deep learning feature extractors integrated with an enhanced ensemble algorithm based on the performance metrics for the imbalanced data. The Support Vector Machine (SVM) classifier is utilized as the last layer of each deep learning component and also as the weak learners in the ensemble module. The framework is evaluated on two large-scale and imbalanced video datasets (namely, disaster and TRECVID). The extensive experimental results illustrate the advantage and effectiveness of the proposed framework. It also demonstrates that the proposed framework outperforms several well-known deep learning methods, as well as the conventional features integrated with different classifiers.

Keywords: Deep learning; convolutional neural networks; ensemble learning; imbalanced data; video event detection; multimedia big data.

1. Introduction

Multimedia data is an inimitable source of information which presents both opportunities and challenges [7, 9, 49, 50, 51, 59]. During the last few years, multimedia sources (e.g., Flickr, Twitter, YouTube, etc.) have produced an extensive amount of data. For example, Twitter active users generate about 500 million tweets everyday or YouTube users upload hundred hours of videos every minute. This explosion of multimedia data including image, audio, video, and text, has turned it into a big data problem.

The necessity of automatic semantic analysis in multimedia data is apparent in many real-world applications [13]. Specifically, video event detection is an important and chal-

lenging task in multimedia management systems. Over the last decade, researchers have been looking for automatic techniques to detect the most interesting events and concepts from videos [15, 35, 38]. Some applications of video semantic event detection include crime detection, natural disaster retrieval, and interesting event recognition from sport games.

In spite of all the opportunities provided by multimedia big data and semantic analysis, several challenges need to be addressed. The first challenge is how to manage such large and multi-modal data in an effective and efficient manner. In addition, imbalanced data (data with a non-uniform or skewed distribution) problem is intricate and conventional straightforward techniques are incapable of handling it [44]. The latter issue has been widely observed in multimedia concept detection applications. Cancer detection from medial data (e.g., MRI images) or fraud detection from bank transactions are examples of imbalanced data. In addition, retrieving videos containing natural disasters among thousands of YouTube videos can be also considered as an imbalanced data problem. In the latter example where meta-data and textual information may not be always reliable and accurate, we are looking for an interesting or minor event (videos containing disaster information); while the distribution of classes (minor and major ones) are non-uniform. The skewed distribution of data in a video event detection task is inevitable and conventional learning techniques are mainly biased toward the major class, while we are interested in the minor one. Although this issue has been studied in the literature for many years [21, 30, 37], it is still a hot topic especially in large-scale multimodal datasets.

Deep learning is an emerging topic which has led to many breakthroughs for multimedia research including image processing, text mining, and speech recognition [54]. It has attracted lots of attentions in industry and academia in the past few years [31, 40, 53]. In general, a deep graph architecture contains a cascade of layers, composed of multiple linear and non-linear transformations. Using this architecture, deep learning models very high-level abstractions from the raw data. However, there is limited work that handles the imbalanced data problem in multimedia data using deep learning.

Till now, numerous deep learning architectures have been proposed for a variety of applications. However, it is almost impossible for a single model to work well for all scenarios and datasets. It is also difficult to handle imbalanced and big multimedia data because of over-fitting, information loss, and additional bias [58]. Therefore, multi-class fusion can be utilized to improve deep learning techniques.

In this study, we propose an ensemble deep learning framework, which not only overcomes the imbalanced data issue in multimedia big data, but also decreases the information loss and over-fitting problems caused by single models. Inspired by the great success of deep learning, it is used for deep feature analysis with the application to video event detection. Thereafter, an ensemble approach is developed based on the performance of each weak learner (Support Vector Machine (SVM) classifier) on each deep feature set to improve the semantic event detection in imbalanced datasets.

This paper is organized as follows. An overview of the state-of-the-art research in imbalanced multimedia analysis is provided in section 2. Section 3 presents the details of the proposed ensemble deep learning framework. In section 4, a comprehensive experimental analysis and its results are discussed. Lastly, the paper is concluded in section 5.

2. Related Work

The proliferation and explosion of multimedia data including audio, image, video, and text has led to emergence of a variety of applications in multimedia management systems [6, 24, 47, 48, 55, 59]. Among them, video analysis is one of the most challenging topics [12, 39]. More specifically, automatic video event detection has attracted lots of attentions in real-world applications [8, 11] such as video surveillance, disaster information management, crime detection, to name a few.

Nevertheless, to fulfill all requirements in a video analysis application, there are several challenges that need to be addressed. First of all, the existence of an interesting semantic concept or event in a video data can be rare. In other words, we are facing an imbalanced data problem. This phenomenon has been widely seen in real-world applications [30], such as activity recognition and video mining [14, 51].

In general, the solutions to imbalanced data can be divided into three categories [22]. The most common group is the resampling methods such as oversampling and undersampling. These techniques mainly adjust the data distribution in order to create a balanced dataset. The second group is the cost-sensitive techniques which penalize the misclassified instances and try to improve the performance of the minority class. Last group includes kernel-based and active learning methods. The main purpose of this group is to utilize more robust classification algorithms to naturally handle imbalanced data. Among them, ensemble learning (can be classified as the cost sensitive methods) has shown significant successes in the recent years [42, 58].

In a multimedia task, features matter. Therefore, how to extract useful features from data in order to improve the final detection results is another substantial challenge in multimedia systems. A multi-modality fusion technique for multimedia semantic retrieval is discussed in [20]. First, the feature space is reduced by calculating the correlation between feature pairs and those features with low correlation toward others are removed. Afterward, a technique called Hidden Coherent Feature Groups (HCFGs) is utilized for feature grouping [57]. Eventually, for each feature group, a classifier is trained and its scores are fused with other classifiers for final event detection. In another work [41], spatial and temporal information from video sequences are integrated in order to generate a new feature representation. In that work, the authors apply the optical flow field and Harris3D corner detector integrated with an ensemble boosting technique which is based on two classification algorithms: sparse representation and hamming distance classifiers.

Deep learning is a major advancement in machine learning and data mining. Although deep learning has a long history in AI [54], it has become a solution for many problems in recent years with the use of the massive amount of computational power and the design of efficient algorithms [32]. Convolutional Neural Networks (CNNs) [33] are one of the most effective algorithms in deep learning. It is developed around 1990 inspired by the Hubel and Wiesel research about animal visual cortex [25]. The CNNs have been widely applied in image processing and improved the traditional feed-forward neural networks [46]. The main difference lies in its local connectivity and weight sharing topology which remarkably constrains the complexity of the networks. In contrast to the traditional neural networks that

are very prone to over-fitting and hard to interpret (black-boxes) [40], new deep learning algorithms are more interpretable due to the strong local modeling of the deep learning architectures. In addition, not only the computational powers have increased vastly during the last decade, but also the novel ideas, algorithms, and deep network architectures in recent few years have led to a big revolution in multimedia applications, mainly in image recognition and object detection. Alexnet [31] is one of the first new deep learning architectures that significantly improved the performance of deep neural networks and classified extremely large datasets (e.g., ImageNet in ILSVRC 2012) into a thousand classes. Later, a novel deep CNN architecture called GoogleNet is proposed in [53] which is experimentally evaluated on classification and detection tasks of ILSVRC 2014. This network is expanded in both depth and width compared to the conventional networks while the computational costs remain constant. Another successful network is called Regions with convolutional Neural Networks (R-CNN) [19]. It includes two subproblems: object detection and image classification. Finally, Microsoft introduced its Residual deep network in 2015 which beats human in object detection in ILSVRC 2015 [23]. After that, no significant improvement in deep learning architectures has been reported except those trying to integrate the advantages of current existing techniques in a reasonable way.

In this paper, we not only extract the beneficial parameters from the aforementioned deep networks, but also develop a new and advanced ensemble technique to integrate the results in an effective manner. The motivation is the fact that a single classifier may not be always capable of managing large and noisy datasets with skewed distributions. However, ensemble algorithms can be used to improve the classification performance by taking advantages of several classifiers [4]. In [56], a scalable ensemble classifier is proposed based on the decision of several "judgers" which are previously trained and evaluated on different classifiers and features. In another work, a positive enhanced ensemble algorithm is proposed [58]. The algorithm handles the imbalanced data problem in video event retrieval by combining a sampling-based technique with a fusion-based classification approach. This method is applied to an imbalanced sport video dataset to detect the interesting events. Ensemble techniques are also applied to several traditional neural networks to alleviate the effects of over-fitting. For instance, an ensemble neural network is presented in [5]. The approach is developed using the bootstrapped sampling approach integrated with conventional neural networks in order to detect rare events in soccer videos.

3. Ensemble Deep Learning Framework

The Ensemble Deep Learning (EDL) framework consists of a mixture of feature extractors using deep learning techniques which are integrated with the proposed ensemble algorithm. The whole framework is divided into three main modules, namely (1) preprocessing, (2) deep feature extraction, and (3) classification (as shown in Figure 1). The classification module also includes the training, validation, and testing steps.

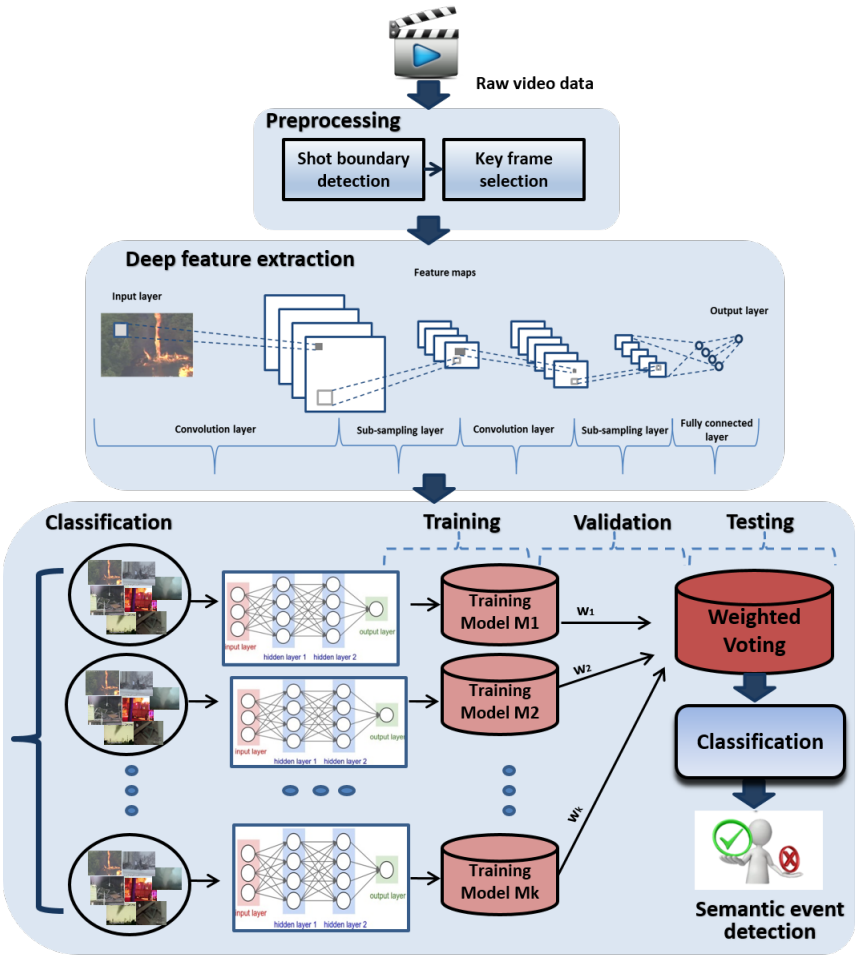


Fig. 1. The proposed ensemble deep learning framework

3.1. *Preprocessing*

The first step in every data analysis is how to preprocess the unstructured data. This step is domain specific and each type of data (e.g., audio, image, video, and text) may require its own preprocessing routines. For video processing in this study, we utilize an automatic and unsupervised shot boundary detection approach [10] based on the object tracking and image segmentation techniques. Using this approach, shot boundaries of each raw video are detected and the first frame of each shot is chosen as the keyframe. The first frame of each shot can be considered as the most distinctive one as it is the boundary of two successive shots. After the preprocessing, the selected keyframes are used for event detection in

videos.

3.2. *Deep feature extraction*

Before 2010, research studies in computer vision mostly focused on improving the handcrafted features and generating more discriminative attributes from the data [34]. Some common and powerful handcrafted features include HOG [16], CEDD [3], and SIFT [43] for visual data and MFCCs [28] for aural data. However, this progress started to slow down between 2010 and 2012 with the advent of new deep learning techniques such as CNNs. In recent years, deep learning is growing very fast and has exceedingly raised the performance results. In this study, we also decide to take advantage of this emerging algorithm and apply it as a feature extractor to our data. For this purpose, several rich and deep feature extraction models are integrated in a proper manner. The deep feature extraction module is based on the CNN algorithm and utilizes the pre-trained models using transfer learning.

3.2.1. *Convolutional neural networks*

CNNs [33] are an advanced version of MultiLayer Perceptron (MLP) networks. However, in CNNs, most neurons are locally connected instead of fully connected, which highly increases the training speed and reduces over-fitting by eliminating a vast amount of parameters in the network.

Unlike MLP, the inputs of each layer in CNNs are arranged in three dimensions: width, height, and depth. For example, for a $256 \times 256 \times 3$ image input, the width and height equal 256 and 3 is the depth of this input which refers to the channel number (e.g., RGB). Each neuron in CNNs is connected to a small region of its previous layer. In overall, there are three main layers to build a convolutional network architecture: (1) Convolutional layer, (2) Pooling layer, and (3) Fully connected layer [1]. A CNN includes a stack of convolutional layers followed by a pooling layer and is usually ended with a fully connected layer as shown in the deep feature extraction module in Figure 1.

In the convolutional layer, the neurons are connected to local regions in the input, each generating a dot product between a small region in the input volume and their corresponding weights. As a result, a number of feature maps are generated, by convolving (sliding) filters over all spatial locations in the input data. In other words, the feature maps are obtained by the convolution of the input data with a linear filter (and bias term addition) followed by a nonlinear activation function as illustrated in Equation (1), where x_{ij}^k refers to the k^{th} feature map at a given layer, i and j are the input dimensions, and x_{ij}^{k-1} is the input data from the previous layer. Filters of the k^{th} layer are determined by W_{ij}^k (weights) and b_j^k (bias). Finally, the activation function or nonlinearity is shown with f. One of the mostly used activation functions for deep learning is Rectified Linear Unit (ReLu) ($f(x) = max(0, x)$) which increases the nonlinearity and shows better performance compared to the conventional ones (e.g., sigmoid, tanh, etc.).

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

After each convolutional layer, there exists a pooling layer which applies a nonlinear downsampling operation along the width and height (spatial dimensions) of the image input given in Equation (2), where β_{ij}^k is a multiplicative bias and $down(.)$ is a subsampling function (e.g., max, average, etc.). Therefore, using the pooling layer, the size of each activation map is reduced, which makes the representation more manageable. It also handles over-fitting and provides additional robustness to the network.

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

Finally, the fully connected layer is used as the last layer of CNNs to compute more high-level reasoning or the class scores. Similar to traditional neural networks, all neurons or activation maps from the previous convolutional-subsampling layer are fully connected to a single neuron in this layer.

3.2.2. *Feature extraction using transfer learning*

In this paper, several advanced and successful deep learning architectures are utilized for visual feature extraction. For this purpose, instead of training an entire CNN from scratch, we take the pre-trained reference models and treat the convolutional networks as feature extractors for new datasets. Theses reference models are pre-trained on very large-scale datasets. Specifically, we select those models which have more impacts on the image processing field in recent years. The ImageNet dataset [17] which contains millions of images with 1000 concept categories is used to train all such models. Therefore, we run these pre-trained models on our datasets and generate the features (also known as CNN codes) for all images. These feature sets are further used for the classification. Table 1 presents a summary of the CNN models used in this paper for feature extraction. As can be seen from the table, a variety of models with different numbers of layers and architectures are used and are further explained below:

- AlexNet [31] can be considered as the first attempt and the most influential one that improved the CNN algorithm for image processing and made it popular again. Basically, this network contains eight layers: five convolutional layers followed by the max pooling and dropout layers (to resist the over-fitting issue), and three fully connected layers. The last fully connected layer generates 1000 possible categories. It also uses data augmentation, including horizontal reflections, image translations, and patch extraction. AlexNet was used to win ILSVRC 2012 and it significantly outperforms the second runner-up (over 10%) in terms of top 5 test error rate.
- CaffeNet [29] was developed by the Berkeley Vision and Learning Center (BVLC). It is the reference model used in the Caffe deep learning software tool which is a replication of AlexNet with certain improvements. For instance, the relighting data-augmentation is not used for training CaffeNet and normalization is applied after the pooling layer. Similar to AlexNet, this model is trained on the ImageNet dataset (for more information, please refer to [2]).

Table 1. Pre-trained reference models for feature extraction

Method	Challenges	# layers	# categories	dataset
AlexNet	ILSVRC 2012	8	1000	ImageNet
R-CNN	ILSVRC 2013 VOC 2012	7	200	ImageNet PASCAL VOC
GoogleNet	ILSVRC 2014	22	1000	ImageNet
ResNet	ILSVRC 2015 COCO 2015	152	1000	ImageNet COCO

- Region based CNN (R-CNN) [19] is another impactful network in computer vision, which is mainly developed for object detection tasks. The process can be divided into two parts: it first generates the region proposals using the bounding box segmentation techniques and then applies a CNN-based classifier to detect objects in those locations. Using this technique, it significantly enhanced the performance results (over 30%) compared to the best results on PASCAL VOC 2012. It is also trained on ImageNet in ILSVRC 2013 in which it generates 200 categories.
- GooglNet [53] is the winner of ILSVRC 2014 in two object detection and classification tasks in which it introduced a deeper and wider network. It can be considered as the first attempt not simply stacking several convolutional and pooling layers in a sequential manner, but having the network pieces happening in parallel. In total, it contains 22 layers of a deep network which remarkably considers memory and power usage by utilizing the extra sparsity of layers. The main piece of this network is known as "Inception" which generates more optimal locality and repeats it spatially.
- Residual network (ResNet) [23] is introduced by Microsoft in ILSVRC 2015 with a great performance (error rate of 3.6%). In addition, it won the COCO detection and segmentation tasks in 2015. This network consists of 152 layers (ultra deep) and leverages the residual connections in the whole network. The main purpose of this network is to address the degradation problem: deeper networks, more saturated accuracy. Therefore, a deep residual network is proposed which allows the network layers to fit a residual mapping.

3.3. *Classification*

After extracting features from the aforementioned reference models using an unsupervised transfer learning, we employ a new ensemble technique to alleviate the over-fitting problem and to improve the performance. First, the extracted deep features are analyzed to find the importance of each feature set extracted from each deep learning model in a supervised manner. In addition, an enhanced ensemble method is proposed to optimally integrate the trained models. This method effectively adjusts the weight coefficients for the classification

module (please refer to Figure 1). First, we train k classification models (weak classifiers in the ensemble), each trained on a feature set. Thereafter, the weight coefficient of each classifier is adjusted based on its classification performance on the validation dataset. The classification step includes two parts: deep ensemble learning and testing.

3.3.1. *Deep ensemble learning*

The training procedure of the proposed deep ensemble learning is illustrated in Algorithm 1. In the first step, we divide the dataset into three parts: training T, validation V, and testing T'. Suppose the training set 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, c_i is the instance class (e.g., for a binary classification task $c_i \in \{0, 1\}$), and N is the size of the training set. Moreover, we store all the feature sets extracted from all deep learning models in Fr. This is another input of the training algorithm.

The proposed ensemble learning can be seen as a bootstrap aggregation (bagging) which involves all the weak learners (classifiers) in the voting. However, in this algorithm, a weighted voting is generated rather than assigning an equal weight to each learner. In addition, the weights are assigned based on a metric for imbalanced data. Therefore, the results are improved toward the minority class, while the performance of the majority class is maintained as high as possible. The weak learners or models are defined as $M = \{M_j,\ j = 1, 2, \cdots, k\}$, each trained using a linear SVM as shown in Lines 1-3 of Algorithm 1, where k is the number of total weak learners. SVM is used as the main classifier as it has shown promising results when it integrates with deep learning [18]. After weak learners M_j ($j = 1, 2, \cdots, k$) are trained using the training instances, each model is evaluated using the validation set V as shown in Lines 4-7 of Algorithm 1. For the evaluation and adjusting the weight coefficients, the F1 measure (Equation (3)) is used which is a number between 0 (the worst case) and 1 (the best case).

$$F1 = 2 * \frac{Precision * Recall}{(Precision + Recall)}; \tag{3}$$

where Precision is the ratio of the positive instances truly predicted (TP) to all positive predicted instances (TP+FP); while Recall is the ratio of the TP to all actual positive instances (TP+FN).

Thereafter, the weight of each trained model M_j is calculated using the ratio of the corresponding F1 score ($F1_j$) to the sum of the scores for all models as shown in Equation (4).

$$W_j = \frac{F1_j}{\sum_{j=1}^{k} F1_j}. \tag{4}$$

The weight coefficient assignees higher values (probability) to the models that are more confident about their prediction. Finally, the algorithm returns each model and its corresponding weight W_j to be further used in the testing module.

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

3.3.2. *Testing*

Algorithm 2 illustrates the testing procedure. The first input of this algorithm includes the testing set $T'\{(t'_i),\ i = 1, 2, \cdots, N_3\}$, where t'_i is the i^{th} testing instance and N_3 is the total number of testing instances. In addition to the instances, for each training model M_j, its feature set F_j and weight matrix W_j are given as the input of this algorithm to predict the labels of the testing instances. In order to achieve this, we calculate a weighted sum of the k models (or weighted voting). As can be seen in Line 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. Afterwards, the final label PL_i is calculated as shown in Line 5 of Algorithm 2. Thus, if the generated weighted sum is greater than half, the label is predicted as positive. Accordingly, the weak learners with higher validation performance have higher impacts on the testing prediction in the proposed ensemble algorithm.

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_js and the weight matrices W_js.
Output: Predicted labels PL_i.

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

4. Experimental Analysis

The proposed Ensemble Deep Learning (EDL) framework can be generally applied to a variety of real-world problems such as image, audio, and text classification. In this paper, we specifically evaluated our framework on two video datasets in order to detect semantic events. The first dataset includes videos containing natural disasters, while the second one is a public large-scale video dataset called TRECVID.

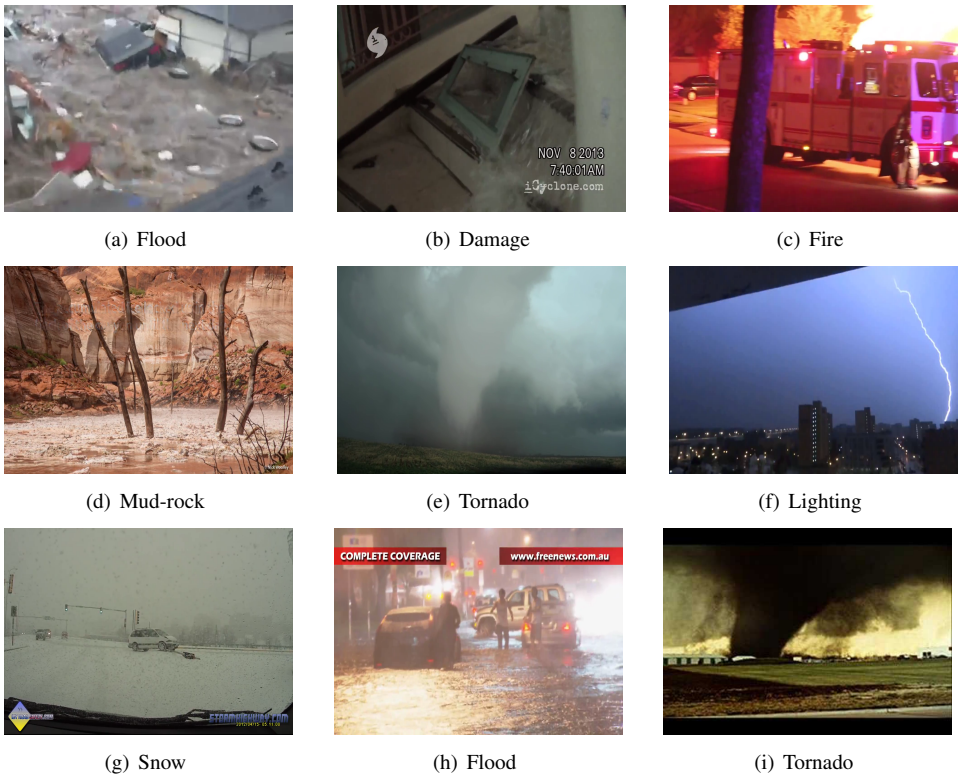
Since both datasets are highly imbalanced, usual metrics such as accuracy and mean-square error may not be effective and reliable. The reason is that the conventional classifiers, which are mostly biased to the majority class, may show very high accuracy on this class while we are more interested in the minority class. Therefore, the proposed framework is evaluated using the common measurement metrics for imbalanced data. Specifically, the confusion matrix parameters including True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), as well as Precision, Recall, and F1 measure (Equation (3)) are employed to evaluate the EDL performance.

The deep learning framework utilized in the following experiments is called Caffe [29]. It includes the advanced deep learning techniques and the state-of-art architectures. The main advantage of Caffe is its rich and updated pre-trained reference models, which can be easily used for fine-tuning and transfer learning. Among all the reference models, we utilize the most successful ones in the literature including R-CNN, CaffeNet, GoogleNet, AlexNet, and ResNet to leverage in the proposed ensemble deep learning framework. To do so, we extract a variety of feature sets from video keyframes using the aforementioned Caffe reference models as explained in Section 3.2.2. All features are extracted from the last fully-connected layer (InnerProduct type) of each model. For instance, layer "fc-rcnn" of R-CNN, layer "fc8" of CaffeNet and AlexNet, "loss3/classifier" of GoogleNet, and "fc1000" of ResNet are used. All models are originally trained on the ImageNet dataset, a very large-scale image database including 1000 classes. The last layer of each selected model generates a 1000-dimension feature vector except the R-CNN, which generates 200 features in its fully connected layer.

4.1. *Evaluation of EDL on disaster dataset*

The disaster dataset is collected from the YouTube videos including seven natural disasters such as flood, damage, fire, mud-rock, tornado, and lightning. In overall, this dataset includes about 80 videos. After applying the video shot boundary detection and keyframe selection techniques, 6884 shots are extracted from this dataset [45]. The average positive/negative (P/N) ratio of the disaster dataset is 0.051. This ratio shows the imbalanced distribution of the data. Figure 2 shows some example keyframes from the disaster dataset.

For the disaster dataset, the proposed EDL framework is compared with two sets of algorithms: the handcrafted features (or engineering features) and the deep learning features. For the first group, several low-level and mid-level features such as HOG, CEDD, color histogram, texture, and wavelet are extracted. The overall feature set for each keyframe includes 707 visual attributes. While, in the second group, the features are generated by applying the deep learning reference models directly on each keyframe. After feature ex-



(a) Flood (b) Damage (c) Fire

(d) Mud-rock (e) Tornado (f) Lighting

(g) Snow (h) Flood (i) Tornado

Fig. 2. Disaster Sample keyframes with annotated concepts

traction, we apply several well-known classifiers including Decision Tree (DT), Multiple Correspondence Analysis (MCA) [36], SVM, and a boosting algorithm on the handcrafted features. The last one is an ensemble algorithm which can be considered as a credible benchmark to be compared with our ensemble framework. The SVM classifier is also utilized for the second group (deep features) as it has been proven to be a successful classifier when it is integrated with deep learning techniques. To have a fair comparison, all the classifiers are tuned to reach to their highest results on this dataset and they are evaluated through a 3-fold cross validation.

Table 2 shows the average precision, recall, and F1 score for both handcrafted and deep feature groups integrated with different classification algorithms. The last row also shows the performance of the proposed EDL algorithm. As can be conclude from the table, the proposed framework improves the performance results in comparison with all the techniques in both groups. In other words, it not only outperforms all the conventional classifiers integrated with the engineering features, but also beats the recent well-known deep neural networks such as GoogleNet and AlexNet. By looking deeper on the results, one can infer that SVM and ensemble (boosting) techniques acquire the highest performance in terms of F1-score in the handcrafted features group. Specifically, SVM has the highest

Table 2. 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**

precision compared to all other algorithms including our proposed EDL. However, its low recall value decreases its overall F1 score. It is worth mentioning that a higher recall value, or in other words lower false negative, is more preferable in an imbalanced data where the correct detection of minority class is vital (e.g., in a cancer detection application). Therefore, integrating the deep features with the SVM classifier in a reasonable manner can increase the recall value and F1 measure significantly as shown in the second group of the results (deep learning features) in Table 2. In this group, AlexNet reaches to the highest F1 score compared to other deep learning techniques. This interesting fact shows that very deep and complex architectures (e.g., GoogleNet) cannot be always useful for different types of datasets. Sometimes a lighter version of deep neural networks not only is more efficient than the complex ones, but also can be generalized for different similar tasks. For example, in this experiment, although the nature of ImageNet is very different with our disaster dataset, but it can be seen that most of the pre-trained models (e.g., AlexNet) on ImageNet can classify disaster events in videos with a reasonable performance. Finally, the proposed framework utilizes the power of deep learning features integrated with a new ensemble technique to improve the overall performance in terms of recall and F1 score. The overall F1 score is calculated as 0.913 which is 4.5% higher than the best classifier in the first group and 2.5% higher than the best result in the second group.

Figure 3 visualizes the performance results of each deep learning algorithm in which the y-axis refers to the F1 score and the x-axis shows different disaster events. It can be seen from this figure that the proposed EDL framework improves the F1 score for all disaster events compared to other techniques. In this figure, in most cases, R-CNN has the lowest performance which can be due to two main reasons. First, its architecture is mainly designed for region-based object detection and semantic segmentation, so that its architecture does not properly match a frame-based video event detection task. In addition, it generates 200 features which include less information than other selected deep learning architectures which generate 1000 features in their last layer. CaffeNet and GoogleNet have achieved very close average performance despite of their different architectures. More specifically, CaffeNet reaches a higher performance for tornado and damage concepts, while GoogleNet

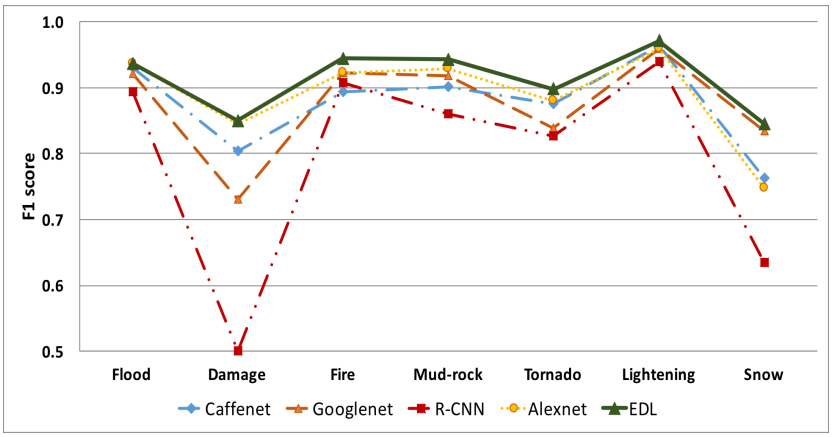


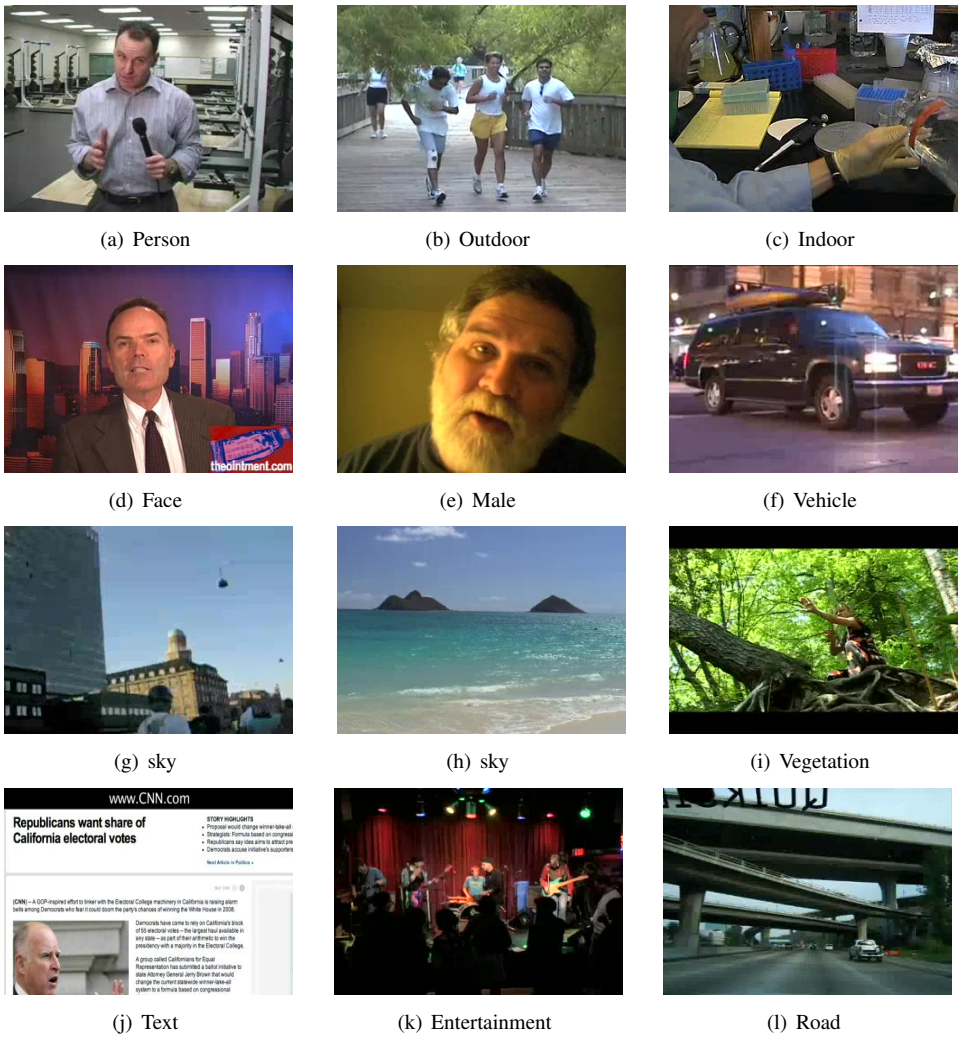
Fig. 3. Performance evaluation for different concepts on the disaster dataset

beats CaffeNet in fire and snow. AlexNet outperforms all other deep learning techniques in terms of F1 score for almost all concepts except lighting and snow. Finally, our proposed technique could successfully improve the results on all semantic events and outperforms the stat-of-the-art deep learning algorithms.

4.2. *Evaluation of EDL on TRECVID dataset*

More experiments are conducted to further demonstrate the effectiveness of the proposed EDL framework. For this purpose, the TRECVID 2011 [52] IACC.1.B dataset including the Internet Archive videos under the Creative Commons licenses is selected as the evaluation benchmark. In the TRECVID semantic indexing (SIN) task, similar to the disaster dataset, the goal is to detect the semantic concepts contained in the video shots. The automatic assignment of semantic labels or tags is a fundamental step for further video browsing, search, and filtering, to name a few. In overall, the IACC.1.B dataset includes hundred thousands of training and testing video keyframes and 346 concepts. In this paper, 20,000 keyframes (the first 10,000 instances from the training and the first 10,000 ones from the testing data) are selected to evaluate our EDL framework. In this dataset, a concept refers to a high-level semantic content or object such as person, vehicle, and sky. Figure 4 demonstrates several sample keyframes in this dataset. The main challenge of this dataset is its highly imbalanced or skewed distribution. Table 3 presents the statistics of the selected concepts in the training and testing sets. These concepts are selected due to their popularity and also the variety of the P/N ratios they have. Therefore, we can evaluate the behavior and functionality of the proposed framework in various situations. For example, the concepts “Person”, “Outdoor”, and “Road” include more positive instances (P/N ratio is above 15%) in the training set compared to other concepts. The average P/N ratios for the training set and testing set are 0.087 and 0.039, respectively.

In this experiment, the results from our framework are compared with the ones from



(a) Person (b) Outdoor (c) Indoor

(d) Face (e) Male (f) Vehicle

(g) sky (h) sky (i) Vegetation

(j) Text (k) Entertainment (l) Road

Fig. 4. TRECVID Sample keyframes with annotated concepts

Tokyo Institute of Technology (TiTech [26, 27]) which was selected as the best performance in the TRECVID 2011 semantic indexing task. The TiTech group extracted several low-level features such as SIFT and MFCCs from each video shot. Therefore, they utilized the handcrafted features including both visual and audio features. Thereafter, an advanced tree-structured Gaussian Mixture Model (GMM) is proposed to model the distribution of low-level features using the maximum a posteriori (MAP) adaptation. In addition, similar to the disaster dataset, the EDL framework is compared with several deep features integrated with the SVM classifier. For this experiment, we replaced the R-CNN with the ResNet due to the low performance of R-CNN in the previous experiment. In addition, the CaffeNet

Table 3. Statistic summary of selected concepts in TRECVID

Concept	Training P/N Ratio	Testing P/N Ratio
Person	0.3714	0.1511
Outdoor	0.1801	0.1490
Indoor	0.0589	0.0218
Face	0.0993	0.0035
Male	0.0418	0.0179
Vehicle	0.0253	0.0560
Sky	0.0275	0.0119
Vegetation	0.0675	0.0233
Text	0.0804	0.0188
Entertainment	0.0413	0.0006
Road	0.165	0.0165

is removed as it has a very similar architecture to the AlexNet. Accordingly, this time we only have three weak learners in our EDL framework.

Table 4 shows the precision and recall values of all deep learning algorithms as well as the ones from the proposed EDL for each concept in this dataset (since the precision and recall values of TiTech framework are not available, they are not listed in this table). As mentioned earlier, the recall metric is more important than the precision in an imbalanced dataset. Thus, the proposed method can achieve the highest recall value (0.436) while maintaining the precision as high as possible as shown in Table 4. Based on this table, the proposed EDL framework beats all other methods in terms of the recall value in most concepts. Although there are a few concepts such as "text" and "'vegetation" having higher recall values in other methods than the EDL, their low precision decreases the overall F1 score significantly. Thus, our proposed method tries to keep the balance between these two metrics and provides higher F1 scores in such concepts. This phenomenon can be also seen in Figure 5, which visualizes the F1 scores for all the benchmark algorithms including the TiTech and all other deep learning frameworks. As can be inferred from the figure, the proposed EDL outperforms all other methods in all concepts except sky. In this concept, although the proposed EDL detects the most positive instances, AlexNet achieves the highest F1 value because of its low false positive or its high precision. Another important fact can be concluded from Figure 5 is the low F1 scores achieved by the TiTech (the best results in the SIN task in TRECVID 2011) compared to deep learning techniques. In other words, similar to the results acquired using the disaster dataset, deep learning features (even those extracted from shallow and simple architectures) contain more information to discriminate the objects than the handcrafted features in many applications.

Finally, Figure 6 depicts the number of positive instances predicted correctly (True Positive) by each deep learning technique. It can be seen that the proposed EDL detects much more positive instances in many concepts such as "person", "outdoor", and "indoor". It also maintains TP as high as possible for highly imbalanced concepts. For instance, the "Enter-

Table 4. Performance evaluation for different concepts on the TRECVID dataset

	AlexNet		**GoogleNet**		**ResNet**		**EDL**	
concept	pre	rec	pre	rec	pre	rec	pre	rec
person	0.321	0.424	0.334	0.487	0.337	0.5	0.332	0.538
outdoor	0.488	0.641	0.502	0.672	0.504	0.681	0.483	0.730
indoor	0.500	0.019	0.333	0.005	0.175	0.052	0.2	0.066
face	0.075	0.514	0.085	0.571	0.085	0.714	0.092	0.571
male	0.529	0.051	0.531	0.097	0.392	0.114	0.393	0.125
vehicle	0.481	0.094	0.494	0.145	0.393	0.158	0.407	0.177
sky	0.294	0.339	0.220	0.373	0.201	0.398	0.208	0.441
vegetation	0.269	0.553	0.303	0.662	0.292	0.754	0.328	0.675
text	0.177	0.595	0.178	0.665	0.178	0.616	0.200	0.643
entertainment	0.083	0.667	0.071	0.167	0.022	0.167	0.083	0.667
road	0.167	0.025	0.397	0.167	0.213	0.16	0.397	0.167
Average	0.308	0.357	0.313	0.365	0.254	0.392	0.284	**0.436**

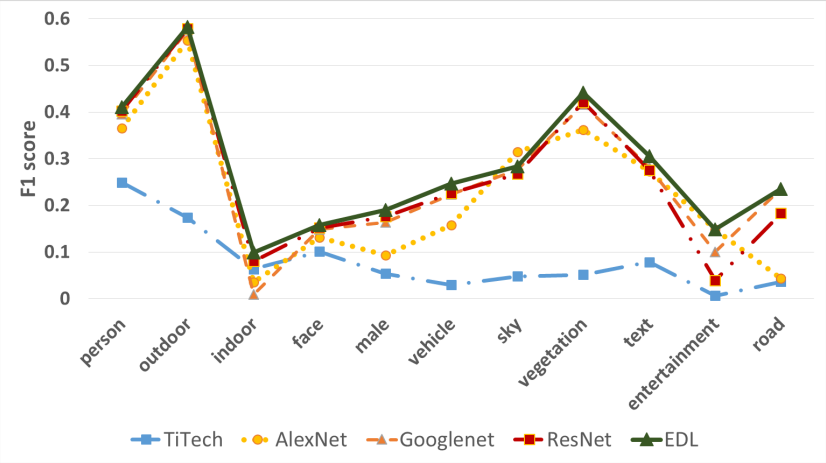


Fig. 5. Performance evaluation for different concepts on the TRECVID dataset

tainment" concept has the lowest TP in the testing set (please refer to Table 3), which means that about only six instances are positive among 10,000 instances in this concept. The EDL and AlexNet can detect four out of six positive instances, while the GoogleNet and ResNet can only detect one positive instance and other five ones are classified as negative.

In summary, based on the experiments on two different datasets with skewed distributions, the proposed EDL achieves promising performance compared to other well-known techniques in this area.

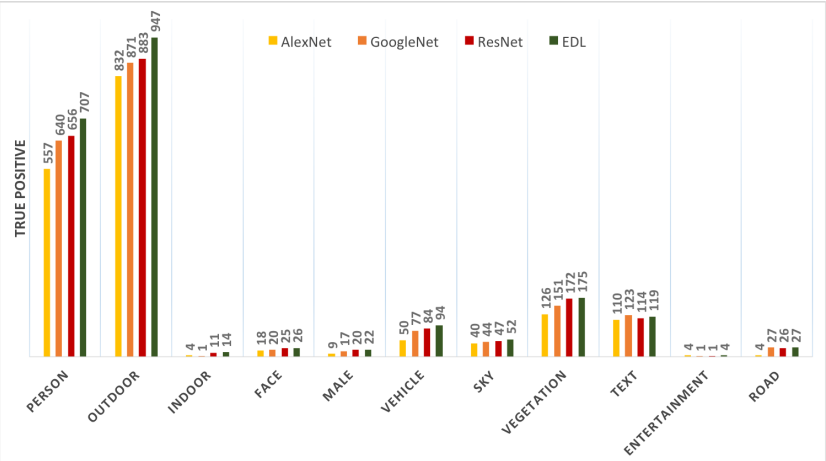


Fig. 6. A comparison of True Positive value for different concepts on the TRECVID dataset

5. Conclusion

As multimedia data is growing exponentially, more advanced techniques are needed to handle such huge amount of unstructured data. In addition, video event detection can be considered as one of the most challenging topics in multimedia management systems. In this paper, we target video semantic event detection with the goal of handling large and imbalanced datasets. We also study the advantages of utilizing deep learning techniques for feature analysis and transfer learning. First, multiple feature sets are extracted from the well-known deep learning algorithms. Thereafter, a novel ensemble deep classifier is developed to fuse different deep feature sets, as well as the results from each weak learner. This framework alleviates the issue of imbalanced data, a very prevalent and unavoidable problem in real-world applications. The proposed framework is extensively evaluated using two large-scale video datasets, namely a natural disaster dataset and the popular TRECVID dataset. The experimental analysis has been conducted to compare the performance of the proposed EDL framework with the ones in other state-of-the-art machine learning algorithms. Specifically, its performance is compared with both handcrafted and deep features groups, integrated with several well-known classifiers. Based on the experimental results, the proposed framework outperforms both groups of algorithms in two datasets with different concepts, which demonstrate its advantage effectiveness for video event detection.

Acknowledgements

For Shu-Ching Chen, this research is partially supported by DHSs VACCINE Center under Award Number 2009-ST- 061-CI0001 and NSF CNS-1461926.

References

[1] J. Bouvrie, Notes on convolutional neural networks, tech. rep., Massachusetts Institute of Technology (2006).

[2] Brewing ImageNet. (2016), `http://caffe.berkeleyvision.org/gathered/examples/imagenet.html`.

[3] 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, (Santorini, Greece, 2008), pp. 312–322.

[4] C. Chen, Q. Zhu, L. Lin and M.-L. Shyu, Web media semantic concept retrieval via tag removal and model fusion, *ACM Transactions on Intelligent Systems and Technology (TIST)* **4**(4), p. 61 (2013).

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

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

[7] S.-C. Chen, M.-L. Shyu and R. Kashyap, Augmented transition network as a semantic model for video data, *International Journal of Networking and Information Systems, Special Issue on Video Data* **3**(1), 9–25 (2000).

[8] S.-C. Chen, M.-L. Shyu, S. Peeta and C. Zhang, Learning-based spatio-temporal vehicle tracking and indexing for transportation multimedia database systems, *IEEE Transactions on Intelligent Transportation Systems* **4**(3), 154–167 (2003).

[9] 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, (CA, USA, 2001), pp. 213–218.

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

[11] S.-C. Chen, M.-L. Shyu, C. Zhang and M. Chen, A multimodal data mining framework for soccer goal detection based on decision tree logic, *International Journal of Computer Applications in Technology* **27**(4), 312–323 (2006).

[12] 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* **10**(04), 715–734 (2001).

[13] S.-C. Chen, S. Sista, M.-L. Shyu and R. L. Kashyap, Augmented transition networks as video browsing models for multimedia databases and multimedia information systems, in *11th IEEE International Conference on Tools with Artificial Intelligence* IEEE, (IL, USA, 1999), pp. 175–182.

[14] 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)* IEEE, (CA, USA, 2005), pp. 8–pp.

[15] 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)* **39**(2), 228–233 (2009).

[16] 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)* **1**, IEEE, (CA, USA, 2005), pp. 886–893.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, (FL, USA, 2009), pp. 248–255.

[18] Z. Ge, C. McCool, C. Sanderson and P. Corke, Content specific feature learning for fine-grained

plant classification, in *Working notes of CLEF 2015 conference* (Toulouse, France, 2015).

[19] 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* (OH, USA, 2014), pp. 580–587.

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

[21] 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, (FL, USA, 2015), pp. 473–487.

[22] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284 (2009).

[23] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (NV, USA, 2016), pp. 770–778.

[24] 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 *Proceedings of the Third International Conference on Multimedia Data Mining* Springer-Verlag, (Alberta, CA, 2002), pp. 100–108.

[25] D. H. Hubel and T. N. Wiesel, Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat, *Journal of neurophysiology* **28**(2), 229–289 (1965).

[26] N. Inoue, Y. Kamishima, T. Wada, K. Shinoda and S. Sato, TokyoTech+Canon at trecvid 2011, *Proc. TRECVID Workshop 2011* (2011).

[27] N. Inoue and K. Shinoda, A fast and accurate video semantic-indexing system using fast MAP adaptation and GMM supervectors, *IEEE Transactions on Multimedia* **14**(4), 1196–1205 (2012).

[28] C. Ittichaichareon, S. Suksri and T. Yingthawornsuk, Speech recognition using MFCC, in *International Conference on Computer Graphics, Simulation and Modeling (ICGSM)* (Pattaya, Thailand, 2012), pp. 28–29.

[29] 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, (FL, USA, 2014), pp. 675–678.

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

[31] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Advances in neural information processing systems* (NV, USA, 2012), pp. 1097–1105.

[32] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature* **521**(7553), 436–444 (2015).

[33] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11), 2278–2324 (1998).

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

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

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

[37] L. Lin, G. Ravitz, M.-L. Shyu and S.-C. Chen, Effective feature space reduction with imbal-

anced data for semantic concept detection, in *IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing (SUTC)* IEEE, (Taichung, Taiwan, 2008), pp. 262–269.

[38] L. Lin and M.-L. Shyu, Weighted association rule mining for video semantic detection, *Methods and Innovations for Multimedia Database Content Management* **1**(1), 37–54 (2012).

[39] L. Lin, M.-L. Shyu, G. Ravitz and S.-C. Chen, Video semantic concept detection via associative classification, in *IEEE International Conference on Multimedia and Expo (ICME)* IEEE, (NY, USA, 2009), pp. 418–421.

[40] M. Lin, Q. Chen and S. Yan, Network in network, *CoRR* **abs/1312.4400** (2013).

[41] 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)* **6**(1), 1–18 (2015).

[42] 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)* **39**(2), 539–550 (2009).

[43] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International journal of computer vision* **60**(2), 91–110 (2004).

[44] T. Meng and M.-L. Shyu, Leveraging concept association network for multimedia rare concept mining and retrieval, in *IEEE International Conference on Multimedia and Expo (ICME)* IEEE, (Melbourne, Australia, 2012), pp. 860–865.

[45] 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, (PA, USA, 2016), pp. 556–564.

[46] S. Pouyanfar and S.-C. Chen, Semantic event detection using ensemble deep learning, in *The IEEE International Symposium on Multimedia (IEEE ISM)* (CA, USA, 2016), pp. 203–208.

[47] S. Pouyanfar and H. Sameti, Music emotion recognition using two level classification, in *Iranian Conference on Intelligent Systems (ICIS)* (Bam, Iran, 2014), pp. 1–6.

[48] M.-L. Shyu, S.-C. Chen and R. L. Kashyap, Generalized affinity-based association rule mining for multimedia database queries, *Knowledge and Information Systems* **3**(3), 319–337 (2001).

[49] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen and N. Zhao, Collaborative filtering by mining association rules from user access sequences, in *International Workshop on Challenges in Web Information Retrieval and Integration (WIRI)* IEEE, (Tokyo, Japan, 2005), pp. 128–135.

[50] M.-L. Shyu, K. Sarinnapakorn, I. Kuruppu-Appuhamilage, S.-C. Chen, L. Chang and T. Goldring, Handling nominal features in anomaly intrusion detection problems, in *15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications (RIDE-SDMA)* IEEE, (Tokyo, Japan, 2005), pp. 55–62.

[51] 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* **10**(2), 252–259 (2008).

[52] A. F. Smeaton, P. Over and W. Kraaij, Evaluation campaigns and trecvid, in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval* ACM, (CA, USA, 2006), pp. 321–330.

[53] 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* (MA, USA, 2015), pp. 1–9.

[54] 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* (FL, USA, 2014), pp. 157–166.

[55] Y. Yan, S. Pouyanfar, H. Tian, S. Guan, H.-Y. Ha, S.-C. Chen, M.-L. Shyu and S. Hamid,

Domain knowledge assisted data processing for florida public hurricane loss model (invited paper), in *IEEE 17th International Conference on Information Reuse and Integration (IRI)* (PA, USA, 2016), pp. 441–447.

[56] Y. Yan, Q. Zhu, M.-L. Shyu and S.-C. Chen, A classifier ensemble framework for multimedia big data classification, in *The 17th IEEE International Conference on Information Reuse and Integration (IRI)* IEEE, (PA, USA, 2016), pp. 615–622.

[57] Y. Yang, Exploring hidden coherent feature groups and temporal semantics for multimedia big data analysis, PhD thesis, Florida International University (FL, USA, 2015).

[58] 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, (CA, USA, 2015), pp. 82–89.

[59] Q. Zhu, L. Lin, M.-L. Shyu and S.-C. Chen, Effective supervised discretization for classification based on correlation maximization, in *IEEE International Conference on Information Reuse and Integration (IRI)* IEEE, (NV, USA, 2011), pp. 390–395.