Enhancing Multimedia Imbalanced Concept Detection Using VIMP in Random Forests

Saad Sadiq[1], Yilin Yan[1], Mei-Ling Shyu[1], Shu-Ching Chen[2], Hemant Ishwaran[3]
[1]*Department of Electrical and Computer Engineering*
University of Miami, Coral Gables, FL, USA
[2]*School of Computing and Information Sciences*
Florida International University, Miami, FL, USA
[3]*Department of Public Health Sciences, Division of Biostatistics*
Miller School of Medicine, University of Miami, Miami, FL, USA
Emails: {saadsadiq@miami.edu, y.yan4@umiami.edu, shyu@miami.edu,
chens@cs.fiu.edu, hishwaran@biostat.med.miami.edu}

Abstract

Recent developments in social media and cloud storage lead to an exponential growth in the amount of multimedia data, which increases the complexity of managing, storing, indexing, and retrieving information from such big data. Many current content-based concept detection approaches lag from successfully bridging the semantic gap. To solve this problem, a multi-stage random forest framework is proposed to generate predictor variables based on multivariate regressions using variable importance (VIMP). By fine tuning the forests and significantly reducing the predictor variables, the concept detection scores are evaluated when the concept of interest is rare and imbalanced, i.e., having little collaboration with other high level concepts. Using classical multivariate statistics, estimating the value of one coordinate using other coordinates standardizes the covariates and it depends upon the variance of the correlations instead of the mean. Thus, conditional dependence on the data being normally distributed is eliminated. Experimental results demonstrate that the proposed framework outperforms those approaches in the comparison in terms of the Mean Average Precision (MAP) values.

Keywords: Multimedia imbalanced concept detection; Multivariate regression; Variable importance (VIMP); Random forests

1. Introduction

The complexity and cost of the data storage and retrieval for multimedia research and applications have increased tremendously [10,14,21,25,26,28,47]. How to store and index multimedia data in various media types including video, audio, image, text, etc. for efficient and effective data retrieval has drawn a lot of attention [16, 31, 42, 43]. To solve this problem, multimedia data is labeled with respect to their real high-level semantic meanings such as "Person", "Boat", and "Football". These labels are often referred to as "concepts" or "semantic concepts" [8, 32, 41, 44]. The foremost challenge in this research domain is to reduce the gap between the low-level features [19, 29] and high-level semantic concepts [7, 10, 15, 29, 48], i.e., to build a connection between the different meanings and conceptions formed by different representation systems.

To bridge the semantic gap [27, 58, 59], a lot of effort has been put into Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) based feature detectors [9, 11–13, 15, 45]. Other methods try to increase the ratio of positive and negative data (for example, video frames) to improve the classification accuracy for automatic labeling and to build the correlations between the labeled concepts to utilize underlying predictors [6, 30, 40, 46, 55, 57]. Some notable solutions include the conditional random field (CRF) methods that improve object classification by maximizing its inter-label agreements [12, 37]. In [34], the CRF method is extended by creating a database of semantic concepts for event detection. On a similar pattern, the ontology based methods utilize the fusion of concept detection confidence scores such as fused Neural Network and concept ontologies to improve the concept identification [4]. In [18], the authors fused the ontologies with fuzzy logic to deduce the correlations among concepts. Other correlation based frameworks such as [24] introduced a Domain Adaptive Semantic Dif-

fusion (DASD) based approach to capture the correlations using Pearson Product. More recent ontology based models use linguistic ontology models to correlate different concepts [2]. For instance, [3, 45] united the WordNet model and Association Rule Mining (ARM) for video retrieval. A more recent and promising approach is to use tree based frameworks that model the contextual correlation using a probabilistic tree method and the conditional probability to evaluate the scores using weights [1, 17]. The bag-of-words (BoW) model in [51] effectively uses random forests and K-Nearest Neighbor (KNN) for large datasets. Similar models assign each descriptor to a single concept or multiple concepts using KNN [36, 52, 56].

Random forests are a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks. Using random forest classifiers, [20] proposed a framework for similarity based labeling of concepts to cluster the training images. It has been observed in [53] that the soft assignment to multiple concepts improves the prediction at the cost of an increased computation time. An interesting framework using random forests and supervised learning reported an improvement in the processing time with a smaller number of classes [35]. An extension of [35] uses random forests in their image segmentation stage by applying the forest on image pixels [39]. However, several random forest based methods reported challenges with noisy attributes and error propagation and their effects on inter-concept collaboration; while others reported shortcomings on either relying on the conditional independence within concepts and depending highly on the prior knowledge and domain knowledge of the data. Some of the data-oriented approaches rely on the assumption that the data is normally distributed and the distribution of the training and testing datasets are the same. These conditions served as the motivation to our work because several of these requirements are not necessarily valid in video dataset detection. Our proposed framework tries to overcome these shortcomings by extending the work from [33,52,56] where the noise issue was minimized and a good retrieval accuracy was achieved by using unsupervised random forests and large datasets.

The paper is organized as follows. In Section 2, the proposed framework is introduced and descriptions are provided for the important components of the developed random forests. Experimental setup based on the TRECVID dataset and the results are discussed in Section 3. Section 4 concludes the paper with the summary of the key findings and important future directions.

2. The Proposed Framework

Our framework is modeled as a random forest based regression problem with big data. The model utilizes the semantic content of images to improve the confidence scores in the retrieval of video shots (keyframes). It was deduced that utilizing the correlations of the concepts assume that the data is normally distributed and centered at zero. This represents a case of conditional expectation and the optimal way to improve the annotation would be to calculate the covariance matrix. However, this is not always the real case so that the proposed model was developed for such cases without the normal distribution assumption. Since there is no "mean" at all, the problem is just a multivariate regression problem with correlation due conditional expectation to calculate the predicted value. This is achieved by using an unsupervised multivariate regression forest that does not require any domain knowledge or does not necessitate any distribution requirement. In classical multivariate statistics, estimating the value of one coordinate using other coordinates standardizes them and the predicted outcome, instead of the mean, depends upon the variance of the correlations.

We consider the scores of 346 concepts from the IACC.1.B dataset in TRECVID 2015 as a 346-dimensional multivariate vector and there are more than 130,000 observations (video shots). Sample images for some of the concepts are depicted in Figure 1.

Figure 1. Sample images of concepts from TRECVID 2015 data

Our proposed framework first splits the TRECVID 2015 data equally into a training data and a testing data. The two data sets are used in the training and testing parts respectively as shown in Figures 2 and 3. The goal is to improve the confidence scores of each concept for all of the observations. Since there is no output variable, we model each instance as a conditional regression problem to predict its best estimate. For any given testing instance, to predict C_i, we take all other variables from C_1, C_2, C_3, ..., C_{i-1}, C_{i+1},

..., C_{346} and regress the value of C_i, using random forests, against this high dimensional large dataset. This process is repeated for all concepts and video shots.

In the training part, a state-of-the-art concept detection framework is applied to the video shots in the training data set and the detection confidence scores for each concept are evaluated. Please note that the focus of this paper is not on the initial concept detection performance but rather on the score improvement in the latter step. Thus, the central part of the proposed framework is kept flexible so that the scores output from any concept detection framework could be utilized with our framework. The variable importance (VIMP) evaluator permutes all 346 concepts and identifies the most significant concepts in the prediction of each concept. This results in significantly reducing the dimensionality and the output of this essential component is used in the testing part. We also grow a synthetic forest to empirically identify the most suitable forest tuning parameters such as mtry and node size for the domain of multimedia concepts detection.

In the testing part, after the detection scores are generated from the concept detection framework, the scores are forwarded to the multivariate regression forest where each concept is predicted as a missing value problem treated by multivariate regression. The VIMP and tuning parameters are used to reduce the dimensionality and fine tune the forest. Finally, the scores output from all the randomly grown trees are assembled together to give the final predicted confidence scores of each concept.

The prediction of each testing video shot is performed by a process called Bootstrap Aggregating (BAGGING). Bootstrap aggregating and random forests were introduced in [5] where it was concluded that the model is always overfitted and by randomly perturbing the dataset and taking the ensemble of that dataset will reduce the overall variance and effectively turn the random forests into highly accurate estimators. It was also proposed that the random forest is a great way for noise reduction and for building a model with low variance [5].

3. VIMP-based Random Forests

3.1. Random Forests

A random forest is an aggregation of *ntree* number of trees, usually in thousands, and each tree is grown by bootstrapping a randomly sampled vector *mtry* from the complete dataset. Each tree in the random forest collection is grown non-deterministically with a two stage method. In the first stage, randomization is induced in each tree by randomly selecting sub-sampled data (bootstrapping) from the original data. The second stage randomization is applied at the node level, where each node is split by randomly selecting a variable from the sub-sampled variables and only those variables are utilized to get the best possible split. This process results in substantially de-correlating the trees so that the final ensemble or the average among the trees will have low variance. Each tree is grown to a depth where the terminal nodes contain at least *nodesize* number of video frames or cases. Algorithm 1 lists the steps of constructing a random forest.

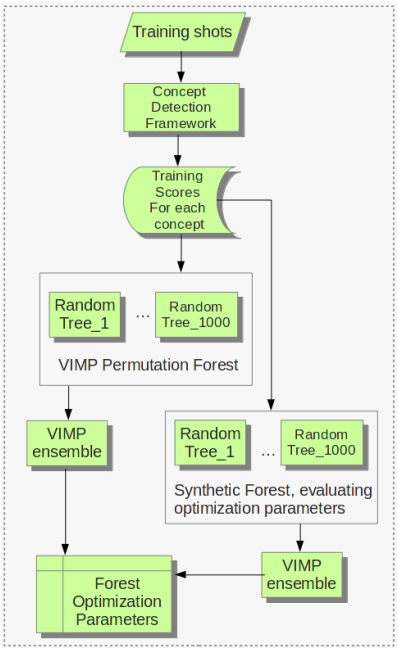

Figure 2. Forest optimization using the training dataset

To achieve this, we begin by modeling the prediction based on the regression setting for which we have a numerical outcome called *Y*. The learned or observed data is assumed to be independently drawn from the joint distribution of (**X**, *Y*) and comprises $n * (p + 1)$ samples, namely $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$. **X** is an n by p matrix indicating the total number of video frames (or samples) and their features *Y*, where **X**=$[\mathbf{x}_1, ..., \mathbf{x}_n]^T$, *Y*=$[y_1, ..., y_n]^T$, $\mathbf{x}_i$ is the subsampled vector (of size 1 by p) from **X** for the i^{th} sample, p is the total number of features (or dimensions), and *Y* indicates the vector of outcome variables (y_i, i=1 to n) that are to be regressed using the random forest.

The random forest for regression is built by growing the trees based on a random vector θ_k such that the tree predictor $h(\mathbf{x}, \theta_k)$ takes on numerical values as opposed to class labels. The vector θ_k contains regressed values of the outcome variable *Y*. The output values are numerical values and we assume that the training dataset is independently drawn from the distribution of the random vector **X** and random

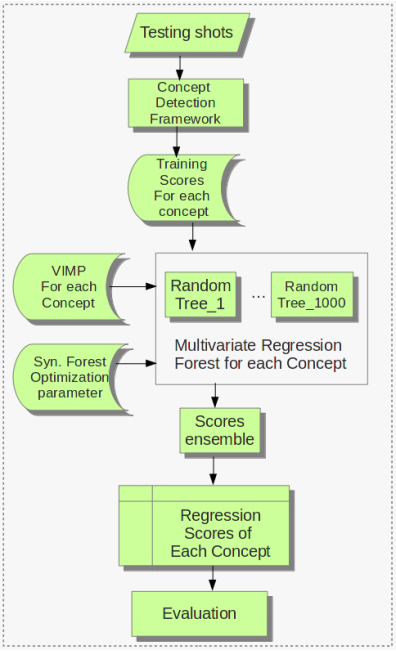

Figure 3. Multivariate regression forest grown on the testing dataset

vector *Y*.

Then, the regression based random forest prediction is defined as the unweighted average over the collection of the predictor trees as shown in Equation (2), where $h(\mathbf{x};\theta_k), k = 1, ..., ntree$ are the collection of the tree predictors and **x** represents the observed input variable vector of length *mtry* with the associated i.i.d random vector θ_k.

$$\overline{h}(\mathbf{x}) = (1/ntree) \sum_{k=1}^{ntree} h(\mathbf{x};\theta_k). \quad (2)$$

As $k \to \infty$, the Law of Large Numbers ensures:

$$E_{\mathbf{X},Y}(Y - \overline{h}(\mathbf{X}))^2 \to E_{\mathbf{X},Y}(Y - E_\theta(\mathbf{X};\theta))^2, \quad (3)$$

where θ represents the regressed outcome variable average over *ntree* trees. The quantity on the right is the prediction (or generalization) error for the random forest, designated PE^*_f. The convergence in Equation (3) implies that the random forests do not overfit. Now the average prediction error for each individual tree is defined in Equation (4).

$$PE^*_t = E_\theta E_{\mathbf{X},Y}(Y - h(\mathbf{X};\theta))^2. \quad (4)$$

The common element in all of these procedures is that for the k^{th} tree, a random vector θ_k is generated, independent of the past random vectors $\theta_1, ..., \theta_{k-1}$ but with the same distribution; and a tree is grown using the training dataset

Algorithm 1 - Construction of Random Forests

1. Draw the *ntree* bootstrap samples from the original data.
2. Grow a tree for each bootstrap data set. At each node of the tree, randomly select *mtry* variables for splitting. Grow the tree so that each terminal node has no fewer than the *nodesize* cases.
3. Aggregate the information from the *ntree* trees for a new data prediction such as majority voting for classification.
4. Compute an out-of-bag (OOB) error rate by using the data not in the bootstrap samples (Equation (1)).

$$MSE_{OOB} = n^{-1} \sum_{i=1}^{n} \{y_i - \widehat{y}_i^{OOB}\}^2, \quad (1)$$

where n indicates the total number of OOB observations (video frames); while y_i and $\widehat{y}_i^{OOB}$ are the average predictions for the in-bag and out-of-bag samples in the i^{th} observation.

and θ_k, resulting in a classifier $h(\mathbf{x},\theta_k)$ where **x** is an input vector. After developing the forest, we further fine tune it by reducing the dimensionality of the features. This is achieved by optimizing *mtry*, *nodesize*, and variable importance (VIMP) as described in the following subsection.

3.2. Optimizing the Forest

There are three key factors to optimize the maximum throughput from a random forest, namely $nodesize$, $mtry$, and VIMP. Their parameters as used in the proposed framework and subsequent justifications are provided as follows. When deciding upon *nodesize*, some methods like [38] argue that large sampled terminal nodes provide consistent results. On the other hand, [5] advises to grow the random forest trees very deeply, i.e., expanding the trees until the terminal nodes contain only one variable. Although this causes very skewed and deep trees that require relatively longer times to compute, it has been observed empirically that near singular terminal sizes are more effective in high dimensional problems [22]. This is because that the trees are grown to purity, i.e., single sampled terminal nodes resulting in a much lower bias. While deep trees result in low bias values, the final ensemble or aggregation of all the trees reduces the variance. Thus we opt our forest to be grown in near purity.

VIMP is another tuning feature of the random forests that we utilize to rank each variable based on its predictability. VIMP calculates the increase in the prediction error for the forest aggregation by randomly noising up a variable and

permuting its value. The larger the VIMP value of each variable, the more predictive the variable is. VIMP helps to select only the most predictive variables in the prediction process and helps implement the dimensionality reduction in an efficient way. Empirical results show that in some cases the number of prediction variables were reduced down to 1%, which also significantly reduced the computation time. The most commonly used permutation method is the Breiman-Cutler importance measure for the random forest. In the method, the variable importance *VI* of a feature variable X_j in tree *k* is evaluated as shown in Equation (5).

$$VI^{(k)}(X_j) = \frac{\sum_{i\in\overline{\mathcal{B}}^{(k)}} I(\gamma_i = \gamma_i^{(k)})}{|\overline{\mathcal{B}}^{(k)}|} - \frac{\sum_{i\in\overline{\mathcal{B}}^{(k)}} I(\gamma_i = \gamma_{i,\pi j}^{(k)})}{|\overline{\mathcal{B}}^{(k)}|}, \quad (5)$$

where X_j is the j^{th} feature from $\mathbf{X}$ and $\overline{\mathcal{B}}^k$ is the out-of-bag (OOB) sample of the variable for a particular tree *k*, with $k \in 1, ..., ntree$. Moreover, $\gamma_i^{(k)}$ is described as the selected class for observation *i* before permuting, $\gamma_{i,\pi j}^{(k)}$ is the class for observation *i* after permuting its value for variable X_j, and $I(.)$ is the identity function. γ_i represents the observed class for the observation *i*. Please note that if variable X_j is not in tree *k*, $VI^{(k)}(X_j) = 0$ by definition. The raw variable importance score for each variable is then computed as the mean importance over all trees as given in Equation (6).

$$VI(X_j) = \frac{\sum_{k=1}^{ntree} VI^{(k)}(x_j)}{ntree}. \quad (6)$$

One of the key techniques in calculating the VIMP variable is to keep the *mtry* variable very close to *p*, where *p* is the total number of predicting variables (in our case 346), and *mtry* is the number of randomly subsampled variables to be used in each tree. The default setting for choosing $mtry$ is $mtry = \sqrt{p}$, but it has been argued in [22] by several empirical studies to keep *mtry* close to $M = 7/8 \times p$. This is because if the *mtry* variables chosen for the root node are noisy (i.e., they are not predictive for the outcome), then the predicted variable and the permuted importance of the variable are also noised up [50]. This principle is depicted in Figure 4, i.e., the larger number of *mtry* helps better identify the variable importance (VIMP). The colors are used to indicate the relevance of the variables with color red being highly predictive.

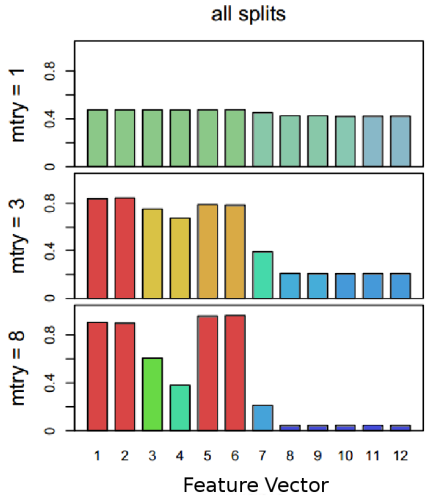

Figure 4. Example of a parametric plot

4. Experiments and Results

4.1 Experiment Setup

For this paper, we use TRECVID 2015 dataset which is a huge dataset with lots of imbalanced concepts. The TRECVID conference series is sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. The goal of the conference series is to encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. The TRECVID dataset is very suitable for our experiment due to its vast volume. We choose the IACC.1.B dataset used in the TRECVID 2015 semantic indexing (SIN) task which aims to detect the semantic concept contained within a video shot. Challenges such as data imbalance [54], scalability, and the semantic gap [27] make the SIN task tough.

In the IACC.1.B dataset, there are 137,327 observations by extracting a keyframe from each shot. Totally 346 concepts are given including many popular semantic concepts include "Vehicle", "Airplane", and "Cloud" which are common and appear in many papers. It also contains many rare and imbalanced concepts such as "Security Checkpoint", "Helicopter Hovering", and "Mosques". The distribution of some concepts are highly skewed in which the majority of the data instances belong to one class and far fewer data instances belong to others. The list of concepts and detailed explanations can be found in [49].

The average precision (AP) value is used as a metric which is widely used in the multimedia concept retrieval domain. For a given concept, $Pre(i)$ indicates the precision at cut-off i in the item list, and N is for the number of the retrieved data instances. The average precision at N (i.e., $AP@N$) is defined in Equation (7). If the denominator is zero, AP is set to zero. By generating AP for all concepts and calculating the mean value of them, the mean average

Table 1. Results Comparison

Framework	MAP@50	MAP@100	MAP@200	MAP@500	MAP@1000
Benchmark	0.1995	0.1895	0.1721	0.1525	0.1362
Naive Bayes	0.1183	0.1203	0.1219	0.1206	0.0728
Random Forests	0.1671	0.1577	0.1523	0.1493	0.1077
VIMP-based Random Forests	**0.2197**	**0.2078**	**0.1881**	**0.1622**	**0.1554**

precision (MAP) value is calculated for evaluation.

$$AP@N = \sum_{i=1}^{N} \frac{Pre(i) \times rel(i)}{\#\ of\ relevant\ instances\ at\ N}; \quad (7)$$

$$rel(i) = \begin{cases} 0, & \text{if instance } i \text{ is negative,} \\ 1, & \text{if instance } i \text{ is positive.} \end{cases}$$

4.2 Experimental Results

In our experiment, we choose 20 highly imbalanced concepts for testing including "Airplane Takeoff", "Emergency Vehicles", "Military", "Natural-Disaster", "US Flags", "Airplane Landing", "Airport Or Airfield", "Car Crash", "Cigar Boats", "Earthquake", "Military Base", "Rowboat", "Election Campaign Debate", "Election Campaign Greeting", "Exiting A Vehicle", "Exiting Car", "Flags", "Military Aircraft", "Rescue Vehicle", and "Prisoner". Also, the detection scores from the group of DVMM Lab of Columbia University [23] for shots are used as the raw scores and the benchmark. Their group got the best performance on TRECVID IACC.1.B dataset but the raw scores for the many imbalanced concepts are relatively low and need to be enhanced.

To conduct the comparison, the proposed framework is evaluated against the following four approaches. The first one, "Benchmark", is the raw scores we got from [23] without any modification. The "Naive Bayes" approach is based on applying the Bayes' theorem with strong independence assumptions between the scores. In the implementation of our approach, the selected 20 imbalanced concepts with the p/n ratio values lower than 0.001 are tested and the VIMP-based random forests are applied. We also compare our work with random forests without VIMP. In the proposed work, the dataset is split in half, one for training and one for testing. The comparison results are shown in Table 1.

As can be seen from Table 1, since the assumption of the "Naive Bayes" approach is not true for many concepts like "sea" and "fish", the accuracy is very low as expected. The random forests without VIMP also fail to enhance the raw scores as well, and this may be caused by the inappropriate tree built process. Among all the four methods, our proposed framework achieves the best performance and successfully enhances the raw scores, which proves the novelty of using random forests with VIMP and shows good MAP results of our proposed framework.

5. Conclusions

Many of the multimedia content based semantic data mining methods face a very complex challenge known as the semantic gap problem. This is the problem of connecting low level details of the image with its high level concepts. The problem becomes even more challenging with those concepts that are rare and imbalanced. In this paper, the proposed framework attempts to solve this problem by utilizing the unsupervised random forest classifiers. Several experiments were conducted on the TRECVID dataset and the results were compared with several existing frameworks. The proposed method illustrates the improvement in terms of the Mean Average Precision (MAP) values for the rare and imbalanced concepts. Furthermore, our proposed random forest approach with VIMP successfully reduces the dependency on domain knowledge and the restriction on data distributions.

Acknowledgment

For Shu-Ching Chen, 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] Y. Aytar, B. O. Orhan, and M. Shah. Improving semantic concept detection and retrieval using contextual estimates. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, pages 536–539. IEEE, 2007.

[2] L. Bai, S. Lao, and J. Guo. Video semantic concept detection using ontology. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 158–163. ACM, 2011.

[3] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video annotation and retrieval using ontologies and rule learning. *IEEE MultiMedia*, 17(4):80–88, 2010.

[4] R. Benmokhtar and B. Huet. An ontology-based evidential framework for video indexing using high-level multimodal

fusion. *Multimedia Tools and Applications*, 73(2):663–689, 2014.

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] L.-C. Chen, J.-W. Hsieh, Y. Yan, and D.-Y. Chen. Vehicle make and model recognition using sparse representation and symmetrical {SURFs}. *Pattern Recognition*, 48(6):1979–1998, 2015.

[7] S.-C. Chen, A. Ghafoor, and R. L. Kashyap. *Semantic Models for Multimedia Database Searching and Browsing*. Kluwer Academic Publishers, Norwell, MA, USA, 2000.

[8] S.-C. Chen and R. Kashyap. Temporal and spatial semantic models for multimedia presentations. In *Proceedings of the International Symposium on Multimedia Information Processing*, pages 441–446, 1997.

[9] S.-C. Chen, S. Rubin, M.-L. Shyu, and C. Zhang. A dynamic user concept pattern learning framework for content-based image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 36(6):772–783, Nov 2006.

[10] 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*, 3(1):9–25, 2000.

[11] S.-C. Chen, M.-L. Shyu, and C. Zhang. An intelligent framework for spatio-temporal vehicle tracking. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, pages 213–218, August 2001.

[12] S.-C. Chen, M.-L. Shyu, and C. Zhang. Innovative shot boundary detection for video indexing. In S. Deb, editor, *Video Data Management and Information Retrieval*, pages 217–236. Idea Group Publishing, 2005.

[13] 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(4):715–734, 2001.

[14] S.-C. Chen, S. Sista, M.-L. Shyu, and R. Kashyap. Augmented transition networks as video browsing models for multimedia databases and multimedia information systems. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence*, pages 175–182, 1999.

[15] 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 *Proceedings of the IEEE International Symposium on Multimedia*, pages 37–44, Dec 2005.

[16] 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.

[17] M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):240–252, 2012.

[18] N. Elleuch, M. Zarka, A. B. Ammar, and A. M. Alimi. A fuzzy ontology: based framework for reasoning in visual video content analysis and indexing. In *Proceedings of the International Workshop on Multimedia Data Mining*, pages 1:1–1:8. ACM, 2011.

[19] J. Fan, H. Luo, and A. K. Elmagarmid. Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. *IEEE Transactions on Image Processing*, 13(7):974–992, 2004.

[20] H. Feng, R. Shi, and T.-S. Chua. A bootstrapping framework for annotating and retrieving www images. In *Proceedings of the ACM International Conference on Multimedia*, pages 960–967. ACM, 2004.

[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 *Proceedings of the International Workshop on Multimedia Data Mining*, pages 100–108, July 2002.

[22] H. Ishwaran, U. B. Kogalur, X. Chen, and A. J. Minn. Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining*, 4(1):115–132, 2011.

[23] Y.-G. Jiang. Prediction scores on TRECVID 2010 data set. http://www.ee.columbia.edu/ln/dvmm/CU-VIREO374/, 2010. Last accessed on September 2011.

[24] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1420–1427. IEEE, 2009.

[25] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht. An effective content-based visual image retrieval system. In *Proceedings of the IEEE International Computer Software and Applications Conference*, pages 914–919, August 2002.

[26] X. Li, S.-C. Chen, M.-L. Shyu, and B. Furht. Image retrieval by color, texture, and spatial information. In *Proceedings of the International Conference on Distributed Multimedia Systems*, pages 152–159, September 2002.

[27] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen. Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 18(3):32–43, March 2011.

[28] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Video semantic concept discovery using multimodal-based association classification. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, pages 859–862, July 2007.

[29] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Effective feature space reduction with imbalanced data for semantic concept detection. In *Proceedings of the IEEE International on Sensor Networks, Ubiquitous, and Trustworthy Computing*, pages 262–269, June 2008.

[30] L. Lin and M.-L. Shyu. Weighted association rule mining for video semantic detection. *International Journal of Multimedia Data Engineering and Management*, 1(1):37–54, 2010.

[31] L. Lin, M.-L. Shyu, G. Ravitz, and S.-C. Chen. Video semantic concept detection via associative classification. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, pages 418–421. IEEE, 2009.

[32] 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*, 6(1):1–18, Jan. 2015.

[33] M. Marszałek, C. Schmid, H. Harzallah, and J. Van De Weijer. Learning object representations for visual object class

recognition. In *Proceedings of the Visual Recognition Challenge Workshop*, 2007.

[34] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, 2012.

[35] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Twentieth Annual Conference on Neural Information Processing Systems*, pages 985–992. MIT Press, 2007.

[36] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[37] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[38] M. R. Segal. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, 2004.

[39] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[40] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang. A unified framework for image database clustering and content-based retrieval. In *Proceedings of the ACM International Workshop on Multimedia Databases*, pages 19–27, New York, NY, USA, 2004. ACM.

[41] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn. Image database retrieval utilizing affinity relationships. In *Proceedings of the ACM International Workshop on Multimedia Databases*, pages 78–85, New York, NY, USA, 2003. ACM.

[42] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak. Mining user access behavior on the www. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 1717–1722. IEEE, 2001.

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

[44] M.-L. Shyu, C. Haruechaiyasak, and S.-C. Chen. Category cluster discovery from distributed www directories. *Information Sciences*, 155(3):181–197, 2003.

[45] M.-L. Shyu, C. Haruechaiyasak, S.-C. Chen, and N. Zhao. Collaborative filtering by mining association rules from user access sequences. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pages 128–135, April 2005.

[46] M.-L. Shyu, T. Quirino, Z. Xie, S.-C. Chen, and L. Chang. Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems*, 2(3):9:1–9:37, 2007.

[47] 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 *Proceedings of the International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications*, pages 55–62. IEEE, 2005.

[48] 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, Feb 2008.

[49] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, October 2006.

[50] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

[51] J. R. Uijlings, A. W. Smeulders, and R. J. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, 12(7):665–681, 2010.

[52] K. E. Van de Sande, T. Gevers, and C. G. Snoek. A comparison of color features for visual concept classification. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*, pages 141–150. ACM, 2008.

[53] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1271–1283, 2010.

[54] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen. Deep learning for imbalanced multimedia data classification. In *Proceedings of the IEEE International Symposium on Multimedia*, pages 483–488, Dec 2015.

[55] Y. Yan, Y. Liu, M.-L. Shyu, and M. Chen. Utilizing concept correlations for effective imbalanced data classification. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*, pages 561–568, Aug 2014.

[56] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[57] 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 *Proceedings of the IEEE International Conference on Semantic Computing*, pages 462–469, 2010.

[58] Q. Zhu, L. Lin, M.-L. Shyu, and D. Liu. Utilizing context information to enhance content-based image classification. *International Journal of Multimedia Data Engineering and Management*, 2(3):34–51, 2011.

[59] Q. Zhu and M.-L. Shyu. Sparse linear integration of content and context modalities for semantic concept retrieval. *IEEE Transactions on Emerging Topics in Computing*, 3(2):152–160, June 2015.