CORRELATION-BASED FEATURE ANALYSIS AND MULTI-MODALITY FUSION FRAMEWORK FOR MULTIMEDIA SEMANTIC RETRIEVAL

Hsin-Yu Ha, Yimin Yang, Fausto C. Fleites, Shu-Ching Chen

School of Computing and Information Sciences,
Florida International University,
1200 SW 8th Street,
Miami, FL 33199, USA
{hha001,yyang010,fflei001,chens}@cs.fiu.edu

ABSTRACT

In this paper, we propose a Correlation based Feature Analysis (CFA) and Multi-Modality Fusion (CFA-MMF) framework for multimedia semantic concept retrieval. The CFA method is able to reduce the feature space and capture the correlation between features, separating the feature set into different feature groups, called Hidden Coherent Feature Groups (HCFGs), based on Maximum Spanning Tree (MaxST) algorithm. A correlation matrix is built upon feature pair correlations, and then a MaxST is constructed based on the correlation matrix. By performing a graph cut procedure on the MaxST, a set of feature groups are obtained, where the intra-group correlation is maximized and the inter-group correlation is minimized. Finally, one classifier is trained for each of the feature groups, and the generated scores from different classifiers are fused for the final retrieval. The proposed framework is effective because it reduces the dimensionality of the feature space. The experimental results on the NUSWIDE-Lite data set demonstrate the effectiveness of the proposed CFA-MMF framework.

Index Terms— multimedia semantic retrieval, feature correlation, maximum spanning tree, multi-modality, fusion

1. INTRODUCTION

Nowadays, the propagation of multimedia data is increasing drastically as a result of advance technology and how people tend to share their life through pictures and videos on a daily basis. This fact has been drawing multimedia research society's attention to implement a comprehensive framework to effectively retrieve a variety of semantic concepts from all kinds of multimedia data, such as images, videos, text, etc.

In order to bridge the semantic gap between the low-level features extracted from multimedia data and their high-level semantic meaning, there are two major challenges researchers have to cope with. First of all, effectively analyzing high-dimensional low-level features in different formats plays an important role in building a good semantic retrieval framework, especially when it comes to scalability issues. To address this issue, researchers usually adopt linear transformations that project the low-level features into a low-dimensional space, reducing the dimensionality of the data as well as the noise contained in the original feature representation. Specifically, statistical measures such as principle component analysis (PCA) and Singular Value Decomposition (SVD) [1, 2] are widely integrated with genetic algorithms (GA) [3, 4] in feature extraction and feature selection to carry out a dimension reduction process. However, projecting all the low-level features into a relatively small universal feature space may be easily affected by outliers and thus valuable information can be lost during dimensionality reduction. Secondly, the correlation between various features and the dependency between modalities should be thoroughly explored since the implication among features would definitely help with semantic retrieval. For example, the tag "sky" implies the color "blue" for the semantic concept "outdoor", which is considered as a "hidden" correlation between features.

Based on the above-mentioned challenges, we propose a Correlation-based Feature Analysis and Multi-Modality Fusion (CFA-MMF) framework for multimedia semantic retrieval. Specifically, we explore the correlations between each feature pair from multiple modalities and the feature space can be reduced by removing features with low correlation toward other features and features with zero standard deviation in the positive instance. Then Maximum Spanning Tree-based Feature Graph Cut (MaxST-FGC) algorithm is used to extract Hidden Coherent Feature Groups (HCFGs), where the intra-group correlation is maximized and the inter-group correlation is minimized. Then one classifier is trained for each of the feature groups, and the generated scores from different classifiers are fused for the final retrieval.

The main contributions of this paper are as follows:

- Propose a correlation-based feature analysis method

that analyzes pair-wise feature relationships and extracts coherent feature groups (i.e., HCFGs) via a graph cut on a correlation-based feature graph.

- Develop a framework that integrates early fusion and late fusion by decomposing all features from multiple modalities into groups, and later fusing the multiple uncorrelated models at the decision level.

The rest of paper is organized as follows. Section 2 introduces the state of the art in multimedia semantic retrieval. Section 3 presents the details of the proposed CFA-MMF framework. Section 4 discusses the experimental results, and section 5 finalizes the paper.

2. RELATED WORK

The related works in the area of multimedia semantic retrieval can be roughly summarized into (1) uni-modality based approaches and (2) multi-modality based approaches, from an information-fusion point of view. In the first category, single modality features, i.e., visual, textual, etc., are extracted for multimedia semantic retrieval. However, due to the versatile characteristics of multimedia data, uni-modality representation cannot properly convey the rich information embedded in the multimedia content. Therefore, many works have been presented for effective fusion of multi-modality features. One common way of multi-modality information fusion is to apply statistical analysis methods to the direct concatenation of features from multiple modalities at feature level. For example, Smaragdis et al. [1] adopt PCA and ICA to obtain the maximally independent audio-video subspaces from the audio-visual concatenated features. Huanzhang et al. [5] apply both PCA and Adaboost as feature selection methodologies to select useful region-based features in object detection. Kusuma et al. [6] exploit the dependency between 2D and 3D facial images and recombined the features from different modalities with the usage of PCA in the first phase. In the second phase, Fishers Linear Discriminant (FLD) was applied to perform another recombination transform into more discriminating data.

Besides analyzing the feature level correlation, the correlation among different models and model confidence toward extracting semantic concepts have also been studied. Liu et al. [7] propose a method called Selective Weighted Late Fusion (SWLF) which used the results trained from a binary classifier to weight the corresponding features in testing data set. Chen et al. [8] propose a fusion strategy to combine ranking scores from both tag-based and content-based models, where the adjustment, reliability, and correlation of ranking scores from different models are all considered. Hofmann et al. [9] propose a fusion method based on probabilistic kernel density estimation to fuse the output of part-based object detectors from multiple camera views in person detection.

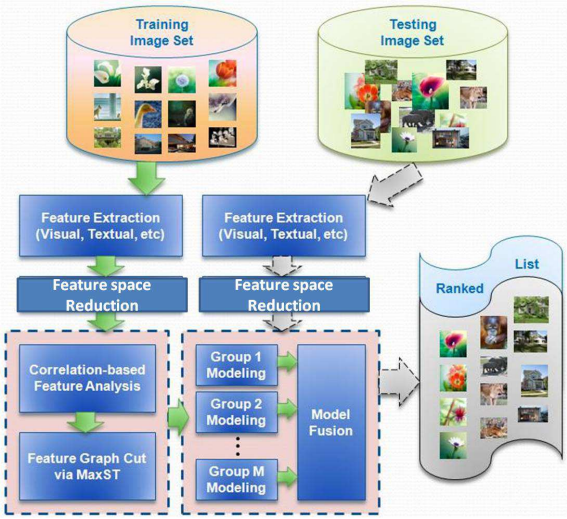

Fig. 1. Correlation based Feature Analysis (CFA) and Multi-Modality Fusion (CFA-MMF) framework.

Despite all the reported advantages reported in existing multi-modality fusion frameworks for multimedia semantic retrieval, they still suffer from the problem of information loss by transformation between different feature spaces and cannot fully utilize the complementary and mutual information among features from multiple modalities. To tackle this problem, we propose the CFA-MMF framework that discovers the HCFGs crossing multiple modalities based on feature correlation analysis and enhance the predictive power of the final fused model.

3. PROPOSED FRAMEWORK

The proposed semantic retrieval framework is depicted in Fig. 1. It builds the retrieval model following a five-step process that consists of (a) feature extraction, (b) pre-processing, (c) correlation-based feature analysis and feature graph cut via MaxST algorithm, (d) model training, and (e) model fusion. Firstly, in the first two steps, the system extracts multi-modality features (e.g., visual, textual, etc.) from the training data and performs pre-processing to normalize the features and remove those with relatively low variance. Secondly, in the correlation-based feature analysis and feature graph cut step, the system computes a feature similarity matrix based on correlation coefficients for all pairs of retained features and applies the MaxST algorithm to analyze the original feature set and obtain HCFGs that exhibit low inter-group correlation and high intra-group correlation. Subsequently, the model training step builds a classification model for each discovered feature group. Finally, the model fusion step combines the individual models using the proposed multi-model fusion strategies. Such a partition of the feature set into HCFGs aims at "untapping" hidden feature groups that

will enhance the predictive power of the fused model. When a query is issued to the system, the system performs feature extraction and pre-processing and groups the features into the same HCFGs identified in the training phase. The HCFGs are then fed to the trained models obtained during the model training step. The generated testing scores are afterward fused and ranked.

3.1. Correlation based Feature Analysis (CFA)

Though the features are extracted from diverse media streams, they may be correlated. For example, in a video shot, the visual frames show a dog barking while the audio channel also records the sound. If the two sources could be effectively integrated in the retrieval system, this kind of multi-modality features may be more discriminant than the single modality feature. On the other hand, the independence among the modalities is also important as it may provide additional cues that help for the retrieval. When fusing multiple modalities, this correlation and independence may equally provide valuable insight based on a particular scenario or context. This section describes the proposed correlation-based analysis method that explores the interrelationship among feature from multiple modalities and constructs the basis for feature graph cut (elaborated in section 3.2).

Suppose a given dataset is denoted by $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^L$ represents each instance in the dataset (N and L are the number of instances and the feature set cardinality, respectively). Then the feature set F is represented as $\{\mathbf{f}^l\}_{l=1}^{L}$, where $\mathbf{f}$ is a feature representation of all the instances in $\mathbf{X}$.

It is worth noting that we do not differentiate the features from multiple modalities and treat each individual feature equally at this point. Let $(\mathbf{f}^j, \mathbf{f}^k)$ $(j, k = 1, 2, \cdots, L)$ be a feature pair, then the correlation coefficient between them can be calculated as follows

$$C_{\mathbf{f}^j,\mathbf{f}^k} = \frac{\sum_{i=1}^{N}(\mathbf{f}_i^j - \overline{\mathbf{f}^j})(\mathbf{f}_i^k - \overline{\mathbf{f}^k})}{\sqrt{\sum_{i=1}^{N}(\mathbf{f}_i^j - \overline{\mathbf{f}^j})^2}\sqrt{\sum_{i=1}^{N}(\mathbf{f}_i^k - \overline{\mathbf{f}^k})^2}}, \qquad (1)$$

where $\overline{\mathbf{f}^j}$ and $\overline{\mathbf{f}^k}$ are the mean values of $\mathbf{f}^j$ and $\mathbf{f}^k$, respectively. The initial feature correlation matrix $\mathbf{C}$ is constructed as

$$\begin{bmatrix} C_{\mathbf{f}^1,\mathbf{f}^1} & C_{\mathbf{f}^1,\mathbf{f}^2} & \cdots & C_{\mathbf{f}^1,\mathbf{f}^L} \\ C_{\mathbf{f}^2,\mathbf{f}^1} & C_{\mathbf{f}^2,\mathbf{f}^2} & \cdots & C_{\mathbf{f}^2,\mathbf{f}^L} \\ \vdots & \vdots & \ddots & \vdots \\ C_{\mathbf{f}^L,\mathbf{f}^1} & C_{\mathbf{f}^L,\mathbf{f}^2} & \cdots & C_{\mathbf{f}^L,\mathbf{f}^L} \end{bmatrix}$$

Each element in the matrix presents the correlation coefficient between each feature pair, and the matrix is symmetric, i.e., $C_{\mathbf{f}^j,\mathbf{f}^k}$ equals $C_{\mathbf{f}^k,\mathbf{f}^j}$

The above correlation coefficients analysis method is based on the calculation of Pearson product-moment correlation coefficient, which assumes normally-distributed data and the linear relationship between feature variables. However, this is not always the case. In order to take into account the situation where the feature variables follow a non-linear relationship, we propose another correlation estimation method based on the Spearman's rank correlation coefficients, which use the ranks of the observations instead of their values and are calculated as

$$C_{\mathbf{r}^j,\mathbf{r}^k} = \frac{\sum_{i=1}^{N}(\mathbf{r}_i^j - \overline{\mathbf{r}^j})(\mathbf{r}_i^k - \overline{\mathbf{r}^k})}{\sqrt{\sum_{i=1}^{N}(\mathbf{r}_i^j - \overline{\mathbf{r}^j})^2}\sqrt{\sum_{i=1}^{N}(\mathbf{r}_i^k - \overline{\mathbf{r}^k})^2}}, \qquad (2)$$

where $\mathbf{r}$ is the rank representation of the feature variable $\mathbf{f}$.

Finally, we applied the following rules to regulate the correlation matrix:

- Only the feature value with non-zero standard deviation from positive instances were considered in obtaining correlation coefficients toward other features.
- The self correlation coefficients are set to zero (i.e., $C_{\mathbf{f}^j,\mathbf{f}^j} = 0$) for the purpose of later feature graph operation.
- The negative correlation coefficients are replaced by their absolute values. This operation is necessary because we are more concerned with how much two features are correlated than how far they depart from each other. In other words, it is not relevant in which direction two features are correlated.
- All the correlation coefficients are calculated based on positive instances. Therefore the correlation matrix is concept specific. This rule implies the advantage of our feature analysis approach by decreasing the total number of training instances and reducing computation complexity, which is a considerable merit over the other statistical-based methods such as PCA, ICA etc.

By using the proposed correlation-based feature analysis method, we are able to capture the correlations among feature variables from multiple multimedia modalities at different granularity. For example, either one the feature ($\mathbf{f}^j$ or $\mathbf{f}^k$) in the correlation coefficient $C_{\mathbf{f}^j,\mathbf{f}^k}$ may be color feature or texture feature from the visual modality, or the tag feature from textual modality, or even the object location feature, which can be considered as a middle level feature based on visual characteristics.

3.2. Feature Graph Cut via Maximum Spanning Tree

To better cope with feature correlation from different modalities, a graph-based approach Maximum Spanning Tree (MaxST) was leveraged in our framework due to its capability of detecting clusters with irregular boundaries. Let $G(F, E)$ be the general notation of a feature graph constructed based

on the feature correlation matrix, where F is the feature set (section 3.1) and E represents the set of feature correlation coefficients $\left\{C_{\mathbf{f}^j,\mathbf{f}^k}\right\}_{j,k=1}^{L}, j < k$. Prim's method [10] was used for constructing a MaxST over the features under absolute correlation value [11]. Unlike other research works using minimum spanning tree to cluster data instance, we constructed an acyclic subgraph that has maximum sum of edge weights and spans over all the vertices. Next, all the edges included in the MaxST are sorted in ascending order. Finally, M feature groups which have high intra-group correlation and low inter-group correlation are obtained by removing $M-1$ smallest edges from the MaxST.

3.3. Model Fusion

The final fusion of the scores from multiple models are based on the refined fusion scheme ARC [8] expressed as

$$Score(I) = \sum_{m=1}^{M} \frac{\xi_m \cdot \theta_m}{\xi_m + \theta_m} \cdot \left(\frac{Score_m(I)}{\omega_m} \right), \qquad (3)$$

where M is the number of models, and ξ_m represents the reliability of model m based on training performance. Specifically, it is calculated as the average precision of the m^{th} model evaluated on the instances in the training set; θ denotes the relationship between the testing score for the m^{th} model and the target concept, which is measured based on the correlation value between the testing score interval and the related concept [8]; ω_m is a scale factor to balance the ranking score for the m^{th} model, which is refined in this paper by using the absolute mean score for all the training instances.

4. EXPERIMENTAL ANALYSIS

In this section, the performance of the proposed CFA-MMF framework is evaluated based on the NUS-WIDE-Lite dataset [12], which includes 55,615 images as well as the associated tags crawled from the Flickr website, with 27,807 for training and 27,808 for testing. The dataset provides the ground truth for 79 concepts and several low-level features commonly used for evaluation such as 64-dimensional color histogram and 128-dimensional wavelet texture, which were also used in this paper. In addition, we also utilize the textual features extracted using the method proposed in [8].

4.1. Experiment Setup

To elaborately evaluate the effectiveness of the presented CFA-MMF framework from different perspectives, we conduct two sets of experiments. First, the CFA and MaxST-FGC algorithms are tested to show the better performance of our proposed feature analysis mechanism against the original flat concatenation of multi-modality features. Second, the overall framework is evaluated to demonstrate the superiority of our proposed approach over the other existing multimedia semantic retrieval works. In both experiments, the correlation-based feature analysis is based on Spearman's rank correlation coefficients and M is selected as 2, i.e., we extract 2 HCFGs from the original feature set. The LibSVM modeling [13] method is adopted in this paper for evaluation because it has been proven to be effective for various multimedia analysis tasks in previous works. It can be easily replaced by any other model training approach. Finally, the evaluation criteria is the well-known Mean Average Precision (MAP) widely used in the information retrieval society.

4.2. Evaluation of CFA and MaxST-FGC algorithms

Fig. 2 shows the number of features after applying the proposed framework (denoted as CFA-MMF) and the original feature set including both visual and textual features (with noisy tag removed [8]). With the proposed correlation-based feature analysis method, only features with high correlation toward other features will be used in classification process. Therefore, there are 11 concepts which had feature dimensionality drop down more than 80% as shown in Fig. 3. As shown in the figure, our framework greatly reduced the dimensionality of the feature space (enhancing the computational performance) and eliminated redundant information.

4.3. Evaluation of CFA-MMF Framework

We compare the results of our proposed framework (CFA-MMF) with other research studies investigating semantic retrieval on the NUS-WIDE-LITE data set using all 79 concepts. These related works demonstrated their performance using K-nearest neighbor (KNN) model [14], LibSVM model [15], linear neighborhood propagation (LNP) [16], entropic graph semi-supervised classification (EGSSC) [17], sparse graph-based semi-supervised learning (SGSSL) [18], large-scale multi-label propagation (LSMP) [19], and three retrieval frameworks, i.e. SVD combined with minimum fusion (SVD+MIN), SVD combined with super kernel fusion (SVD+SKF), and multiple correspondence analysis-based tag removal algorithm (MCA-TR+ARC) constructed from [8].

In our framework, both visual features and image annotated tags are considered as discriminant features to overcome the semantic gap problem . In addition, we apply the MCA-TR method to remove noisy tag information using MCA [8]. Each tag feature was assigned a feature weight and the threshold with highest MAP in the training dataset was set up to remove useless tag features. Other algorithms have their parameters set up which were already proved to be best tuned in [19, 8].

The MAP value of our proposed framework against other above-mentioned frameworks is shown in Fig. 4. and it can

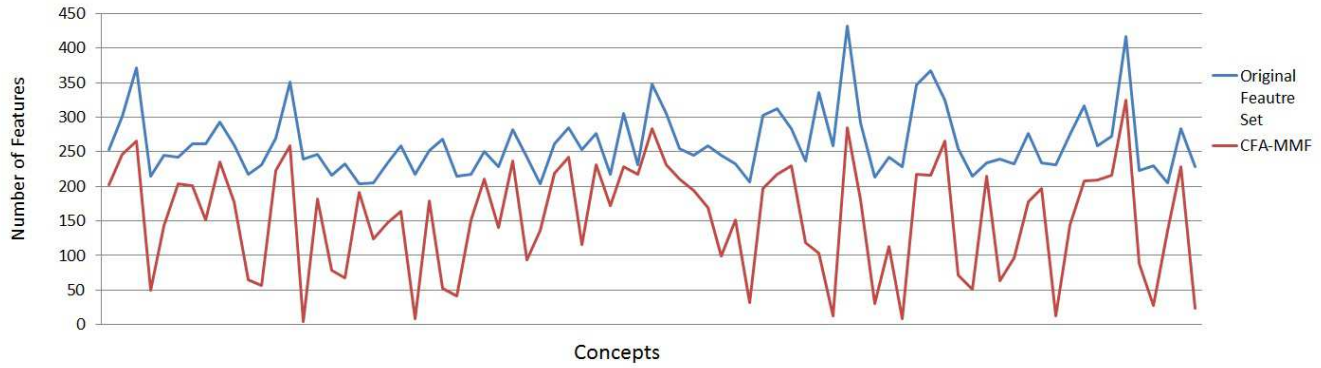

Fig. 2. Dimensionality Reduction of CFA-MMF Over Original Feature Set

Concept	Reduction Rate	Concept	Reduction Rate	Concept	Reduction Rate	Concept	Reduction Rate	Concept	Reduction Rate	Concept	Reduction Rate	Concept	Reduction Rate
airport	-20.16%	flags	**-96.31%**	protest	-59.59%	***tower***	-35.87%	clouds	-26.21%	***military***	-16.61%	surf	-71.65%
animal	-18.27%	flowers	-28.97%	***railroad***	-35.19%	town	-15.81%	computer	**-97.92%**	moon	-20.74%	sunset	-18.46%
beach	-28.49%	food	**-80.22%**	rainbow	**-84.54%**	toy	**-94.37%**	coral	-26.02%	mountain	-25.25%	tiger	-58.80%
bear	-76.74%	fox	**-80.47%**	reflection	-34.77%	***train***	-47.64%	cow	-63.89%	nighttime	-6.06%	zebra	**-89.91%**
birds	-41.63%	frost	-30.28%	***road***	-30.45%	tree	-34.38%	dog	-70.82%	ocean	-18.68%	sports	-53.31%
boats	-15.70%	garden	-16.00%	rocks	-19.01%	valley	-18.99%	***earthquake***	-5.88%	person	-24.51%	leaf	-54.15%
book	-23.28%	glacier	-38.43%	running	-50.21%	***vehicle***	-20.59%	elk	-39.51%	***plane***	-17.65%	cityscape	-17.41%
bridge	-42.37%	grass	-15.96%	***sand***	-69.05%	water	-22.06%	***fire***	-37.18%	plants	-20.82%		
buildings	-19.45%	harbor	-61.16%	sign	**-94.96%**	waterfall	-60.54%	fish	-36.43%	***police***	-34.75%		
cars	-31.54%	horses	-32.84%	sky	-34.03%	wedding	**-87.83%**	swimmers	-76.28%	***statue***	**-96.05%**		
castle	-70.18%	house	-16.09%	***snow***	-38.14%	whales	-33.17%	tattoo	-8.12%	***street***	-37.28%		
cat	-75.76%	lake	-15.09%	soccer	**-85.51%**	window	-19.72%	***temple***	-73.75%	sun	-41.14%		

Fig. 3. Percentage Change in Dimensionality Reduction of CFA-MMF over Original Feature Set

be easily distinguished with at least 4% and at most 30% higher MAP values. Compared with other algorithms, the improvement in performance can be explained as follows. We take advantage of copious information provided along with the image data, which includes features from multiple modalities, and explore the correlation among different modalities to extract a reduced feature set that filters out irrelevant information and identifies feature groups that better fuse information from different modalities.

5. CONCLUSION

In this paper, we have presented a novel correlation-based feature analysis and multi-modality fusion framework for multimedia semantic retrieval. The proposed framework explores the mutual information from multiple modalities by performing correlation analysis for each feature pair and reducing the original feature space. Consequently, the original feature set was separated into HCFGs by using the maximum spanning tree-based feature graph cut algorithm at the feature level. Then a refined multi-modality strategy is employed to combine the testing scores from different training model to obtain optimal performance. The experimental analysis and results demonstrate the effectiveness of the proposed framework. In the future, we would explore more sophisticated correlation analysis method for analyzing versatile feature types and design an adaptive framework to separate the features into multiple feature groups instead of two.

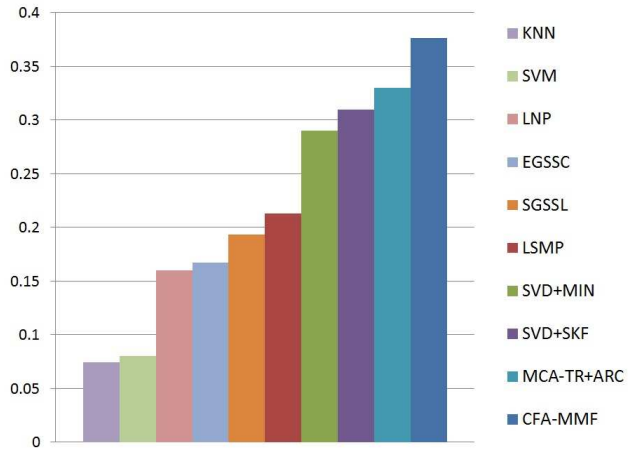

Fig. 4. MAP values of all 79 concepts of the proposed framework and other works on the NUS-WIDE-LITE dataset.

Acknowledgment

This research was supported in part by the U.S. Department of Homeland Security under grant Award Number 2010- ST-062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, and NSF HRD-0833093.

6. REFERENCES

[1] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 709–714.

[2] F. Xiao, M. Zhou, and G. Geng, "Linear transformation technology for image feature drop dimension," in *2011 Fourth International Symposium on Knowledge Acquisition and Modeling (KAM)*. IEEE, 2011, pp. 331–333.

[3] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.

[4] I. Ahmad, A. Abdullah, A. Alghamdi, M. Hussain, and K. Nafjan, "Features subset selection for network intrusion detection mechanism using genetic eigen vectors," in *Proceedings of 2011 International Conference on Telecommunication Technology and Applications (ICTTA 2011)*, 2011, pp. 75–79.

[5] H. Fu, A. Pujol, E. Dellandrea, and L. Chen, "Visual object categorization based on the fusion of region and local features," 2010.

[6] G. Kusuma, C.-S. Chua, and H. Toh, "Recombination of 2d and 3d images for multimodal 2d+ 3d face recognition," in *2010 Fourth Pacific-Rim Symposium on Image and Video Technology (PSIVT)*. IEEE, 2010, pp. 76–81.

[7] N. Liu, E. Dellandrea, C. Zhu, C.-E Bichot, and L. Chen, "A selective weighted late fusion for visual concept recognition," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 426–435.

[8] 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)*.

[9] M. Hofmann, M. Kiechle, and G. Rigoll, "Late fusion for person detection in camera networks," in *2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2011, pp. 41–46.

[10] R.C. Prim, "Shortest connection networks and some generalizations," *Bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[11] P.K. Agarwal, J. Matoušek, and S. Suri, "Farthest neighbors, maximum spanning trees and related problems in higher dimensions," *Computational Geometry*, vol. 1, no. 4, pp. 189–201, 1992.

[12] T.S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, p. 48.

[13] C.C. Chang and C.J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, Pattern Classification and Scene Analysis: Pattern Classification. Wiley, 2001.

[15] I.H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2005.

[16] F. Wang and C. Zhang, "Label propagation through linear neighborhoods," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 985–992.

[17] A. Subramanya and J. Bilmes, "Entropic graph regularization in non-parametric semi-supervised classification," in *Proceedings of Neural Information Processing Society (NIPS)*, 2009, pp. 1803–1811.

[18] J. Tang, S. Yan, R. Hong, G.J. Qi, and T.S. Chua, "Inferring semantic concepts from community-contributed images and noisy tags," in *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 2009, pp. 223–232.

[19] X. Chen, Y. Mu, S. Yan, and T.S. Chua, "Efficient large-scale image annotation by probabilistic collaborative multi-label propagation," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 35–44.