IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 28 September 2014; revised 11 December 2014; accepted 14 December 2014. Date of publication 21 December, 2014;
date of current version 10 June, 2015.

Digital Object Identifier 10.1109/TETC.2014.2384992

Sparse Linear Integration of Content and Context Modalities for Semantic Concept Retrieval

QIUSHA ZHU AND MEI-LING SHYU

Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124 USA

CORRESPONDING AUTHOR: M.-L. SHYU (shyu@miami.edu)

ABSTRACT The semantic gap between low-level visual features and high-level semantics is a well-known challenge in content-based multimedia information retrieval. With the rapid popularization of social media, which allows users to assign tags to describe images and videos, attention is naturally drawn to take advantage of these metadata in order to bridge the semantic gap. This paper proposes a sparse linear integration (SLI) model that focuses on integrating visual content and its associated metadata, which are referred to as the content and the context modalities, respectively, for semantic concept retrieval. An optimization problem is formulated to approximate an instance using a sparse linear combination of other instances and minimize the difference between them. The prediction score of a concept for a test instance measures how well it can be reconstructed by the positive instances of that concept. Two benchmark image data sets and their associated tags are used to evaluate the SLI model. Experimental results show promising performance by comparing with the approaches based on a single modality and approaches based on popular fusion methods.

INDEX TERMS Semantic concept retrieval, sparse linear methods, multimodal integration.

I. INTRODUCTION

Living in a world where digital photo-capture devices has become ubiquitous, more and more people started to share their lives on social networking websites, like YouTube, Flickr, and Facebook. These media repositories allow users to upload images and videos, and edit their metadata, such as titles, descriptions and tags. This new trend has brought a shift in the research of multimedia information retrieval from traditional text-based retrieval to content-based retrieval, and now to a paradigm that needs to integrate both.

Traditional text-based approaches can be traced back to 1970s, which usually relied on manual annotation to perform retrieval. The construction of an index (or a thesaurus) was mostly carried out by specialists, who manually assigned a limited number of keywords to describe the image and video content. Shortly, the processing speed failed to meet the requirements of fast and automatic searches of multimedia content since a manual analysis of multimedia data can be very expensive or simply not feasible when the time is limited or when the amount of data is enormous. In order to organize the vast amount of increasing online multimedia data, learning techniques focused on content analysis have gained popularity over traditional text-based analysis [1]. Content-based approaches were introduced in the early 1990s to classify and retrieve images and videos on the basis of low-level and mid-level visual features. These features are attributes that describe an instance or item, based on color, texture and shape information [2]. Although significant improvements have been achieved by using low-level visual features, the semantic gap challenge still remains. It refers to the difference between high-level semantic concepts (e.g., sky, buildings, dogs, etc.) and extracted low-level visual features (e.g., color, shape, texture, etc.). It is produced by "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [3]. Various advanced features have been detected in order to visually capture the middle-level to high-level semantics contained in an image or a video. However, the semantic gap still exists, and adding more features could also lead to the "curse of dimensionality" [4].

In light of the advantages and disadvantages of both content-based and text-based approaches, studies in recent years have started to investigate how to utilize both approaches to enhance each other [5]. The fundamental property that differentiates these two approaches is the way in which the information is presented, also known as information modality. For content-based approaches, the information is presented by images or videos themselves, which is referred to as the content modality; while for text-based approaches, the information contained in images or videos is presented by texts in the form of metadata, such as titles, descriptions, tags, and surrounding texts. Thus it is referred to as the context modality. On one hand, visual features extracted from the content modality suffer from the semantic gap problem as mentioned before, but they make the automation of organizing multimedia content possible, which can greatly save human efforts of manual annotation. On the other hand, textual features extracted from the context modality usually express the semantics contained in an image or a video, and therefore can bridge the semantic gap that exists in the content modality. However, this metadata is contributed by users, which is known to be imprecise, subjective and uncontrolled. It is too noisy to be used directly as keywords to describe the content. Figure 1 shows some sample images from Flickr together with the user assigned tags. As can be seen, useful tags (tags describing the image content correctly) are embedded in noisy ones.

FIGURE 1. Example social Web images with noisy tags.

Motivated by the complementary information contained in the content and context modalities, a sparse linear integration (SLI) model is proposed in this paper to bridge the semantic gap in content-based semantic concept retrieval by incorporating the associated context information. The integration process is formulated into an optimization problem that aims to minimize the difference between the feature representation of an instance and its reconstructed representation by a sparse linear combination of other instances. The learned model can be directly used for unsupervised applications, which usually involve finding the similarity between two instances [6]. Classification can be performed by using the positive training instances of a concept/class to reconstruct a test instance, and the smaller the reconstruction error is, the more likely that the test instance belongs to this concept/class. The contribution of SLI can be summarized into three folds:

- The fusion of information from different modalities is formulated to the elastic net, which is a regularized regression approach that balances between the lasso using the ℓ_1-norm and ridge regression using the ℓ_2-norm.
- The weights of different modalities are introduced in SLI and the proposed weighting strategy is incorporated in the optimization process through the Lagrange multiplier.
- Each instance is represented by a sparse linear combination of other instances through SLI and a classifier for semantic concept retrieval is developed accordingly based on how well these representations are.

The rest of paper is organized as follows. Section 2 discusses the related work. The detailed problem formalization and solution are presented in Section 3 followed by the experimental results in Section 4. Finally, conclusion is drawn in Section 5.

II. RELATED WORK

In the field of multimedia retrieval, information from different modalities have been utilized to complement each other and have shown promising results in tasks such as semantic concept detection, speech recognition, and multisensor fusion [7]–[9]. Current methods in information fusion typically fall into one of the four branches:

1) Early fusion typically concatenates features from different modalities and results in a single feature representation to be used as input to a learner. This approach is simple and generic but is subject to the "curse of dimensionality" since the concatenated features can easily reach to very high dimensions.
2) Late fusion applies a separate learner to each modality and fuses their decisions through a combiner. Compared to early fusion, late fusion offers scalability and freedom to choose a suitable learning method for each modality. However, it cannot utilize the feature-level correlations from different modalities and is required to make local decisions first.
3) Hybrid fusion involves both early fusion and late fusion by applying early fusion on some modalities and late fusion on the rest of the modalities. Then these decisions are combined in a late fusion manner. Although it offers the flexibility of choosing the proper fusion approach on a subset of modalities, its structure is often application dependent and thus requires domain knowledge.

4) Intermediate fusion is an emerging branch, which does not alter the input feature representation nor require local decisions. It integrates multiple modalities by inferencing a joint model for decision, which often yields superior prediction accuracy [10].

A comparison between early fusion and late fusion was done by Snoek et al. [11], and experiments on broadcast videos for video semantic concept detection showed that late fusion tends to slightly outperform early fusion for most concepts, but for those concepts where early fusion performed better, the gain was more significant. Nagel et al. [12] presented the participation of the Fraunhofer IDMT in the ImageCLEF 2011 Photo Annotation Task. The text-based features were extracted by computing tf-idf values of each tag and visual features were RGB-SIFT descriptors using the codebook approach. In early fusion, both visual and text-based features were considered simultaneously to train the SVM classifier; while in late fusion, two SVMs were trained for each modality and then the results were combined using the geometric mean. The Mean Average Precision (MAP) of 99 concepts showed that the late fusion approach outperformed the early fusion by a very small margin, about 1.5%. An advanced framework proposed in Caicedo et al. [13] connected two data modalities using matrix factorization to project these two matrices into a latent space. Therefore, each representation could be backprojected to the space of the other representation through the common latent space. Then the two backprojected representations were concatenated as well with a weight parameter. Experiments on Corel 5K and MIRFLICKR datasets showed the effectiveness of this framework by comparing with Joint Factorization [14] and their previous work based on Non-negative Matrix Factorization (NMF) [15].

Recently Multiple Kernel Learning (MKL) [16] has been introduced to the domain of heterogeneous feature fusion. It is regarded as intermediate fusion as compared to early fusion and late fusion since kernels are combined as a way to integrate multiple representations. Yu et al. [17] applied MKL to biomedical data fusion. ℓ_2-norm was adopted to get non-sparse optimal kernel coefficients, which was believed to have more advantages over the sparse solution resulted from ℓ_1-norm in real biomedical applications. Yeh et al. [18] proposed a novel multiple kernel learning (MKL) algorithm with a group lasso regularizer for heterogeneous feature fusion and variable selection. It offered a robust way of fitting data extracted from different feature domains by assigning a group of base kernels for each feature representation in an MKL framework. A mixed ℓ_1-norm and ℓ_2-norm constraint enforced the sparsity at the group/feature level and automatically learned a compact feature representation for recognition purposes. Zitnik et al. [10] compared matrix factorization with the state-of-the-art MKL in handling heterogeneous data fusion. A penalized matrix tri-factorization revealed data hidden associations, which simultaneously factorized data matrices. Good accuracy and time response were reported about this new data fusion algorithm.

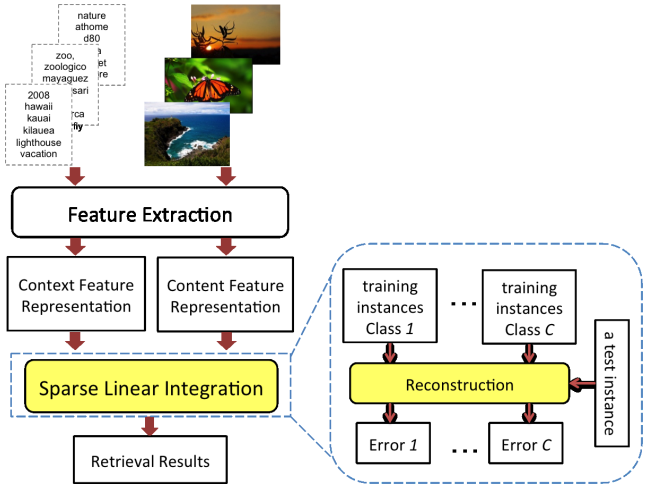

FIGURE 2. **The framework of the SLI Model.**

III. THE PROPOSED SLI MODEL

Compared to early fusion and late fusion, the proposed SLI model is an intermediate fusion approach, which does not generate a combined feature representation, nor does it require local decisions first. The framework of the SLI model is shown in Figure 2. Features are first extracted from each modality to form the content and the context feature representations. In the SLI model, an optimization problem is formulated to approximate an instance using a sparse linear combination of other instances. Given a test instance, a set of sparse coefficients are learned to reconstruct it using the positive training instances belong to a particular Class c, where $c \in \mathbb{R}^C$. The prediction score of the test instance belongs to Class c is measured by Error c, which is the difference between the feature representation of this test instance and its reconstructed representation by the positive training instances belonging Class c weighted by the learned sparse coefficients. The smaller Error c is, the higher probability that the test instance belongs to Class c.

A sparse linear method was first introduced in [19] for top-n recommendation, which generated recommendation results by aggregating user purchase or rating profiles. Later the authors extended the method to incorporate item content information [20], but the basic model was the same. Experiments on various datasets demonstrated high quality recommendations, and the sparsity of the coefficient matrix allowed the method to generate recommendations very fast. Inspired by this method, the SLI model combines two modalities at the intermediate level. A pairwise instance similarity matrix is learned, which can be viewed as the coefficient matrix in matrix factorization [21]. However, instead of factorizing the original feature representation into the basis matrix and the coefficient matrix of a low dimension in the latent space, the original feature representation is used in SLI as the basis matrix. The advantage of SLI is that there is no need to tune the dimension of the latent space as typically required in matrix factorization. This further prevents the potential information loss due to the low-rank approximation.

In addition, we incorporates classification into the integration process, which tends to achieve a higher accuracy compared to methods adopting the early fusion and late fusion approaches that separate integration from classification.

To express and formulate the model in a clear way, the feature representation is denoted by a matrix with each feature or attribute as a row and each instance or item as a column, which is referred to the instance-feature matrix. Thus the feature representation of the content and context modalities can be described by two feature-instance matrix $\boldsymbol{X}^t \in \mathbb{R}^{M^t \times N}$ and $\boldsymbol{X}^v \in \mathbb{R}^{M^v \times N}$, respectively. N is the total number of instances, M^t is the number of features of the context modality and M^v is the number of features of the content modality. Based on the assumption that two instances would have a high correlation if they have similar textual representations as well as similar visual representations, and their correlation would be impaired if they are only similar in textual space or visual space or neither, the correlation coefficient between two instances can be learned by integrating the information from $\boldsymbol{X}^t$ and $\boldsymbol{X}^v$. SLI achieves this by updating the feature representation using a linear combination of the original feature representation weighted by the pairwise instance correlation coefficients. In order to get the updated feature representation, the goal is to learn a pairwise instance coefficient matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ in the updating process, as illustrated in Figure 3. Each column $\boldsymbol{s}_j$ of $\boldsymbol{S}$ denotes the sparse coefficients between instance j, where $j \in \mathbb{R}^N$, and the rest of the instances. Take the context representation $\boldsymbol{X}^t$ for example, Equation (1) expresses that the value of the i-th feature of the j-th instance is updated as a linear combination of the i-th feature of all the instances ($(\boldsymbol{x}_i^t)^T \in \mathbb{R}^N$) and the coefficients between the j-th instance and the other instances ($\boldsymbol{s}_j \in \mathbb{R}^N$). If the j-th instance has a high correlation with the h-th instance, then the i-th feature of the h-th instance would contribute more to the updated value of the i-th feature of the j-th instance, and vice versa. Correspondingly, the update of the j-th column of $\boldsymbol{X}$ is as expressed in Equation (2). The same update applies to $\boldsymbol{x}_j^v$.

$$\boldsymbol{x}_{ij}^t \leftarrow (\boldsymbol{x}_i^t)^T \boldsymbol{s}_j \tag{1}$$

$$\boldsymbol{x}_j^t \leftarrow \boldsymbol{X}^t \boldsymbol{s}_j. \tag{2}$$

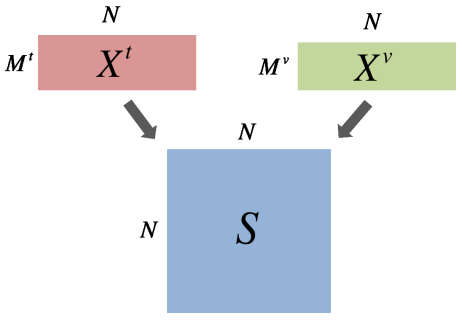

FIGURE 3. The matrix illustration of SLI.

Hence, the problem can be formulated into an optimization problem presented in Equation (3). The terms $\|\boldsymbol{x}_j^t - \boldsymbol{X}^t \boldsymbol{s}_j\|_F^2$ and $\|\boldsymbol{x}_j^v - \boldsymbol{X}^v \boldsymbol{s}_j\|_F^2$ measure how well the update fits $\boldsymbol{x}_j^t$ and $\boldsymbol{x}_j^v$, and α_j^t and α_j^v are their associated weights. Compared to the simple feature concatenation adopted by most of the early fusion approaches, these weights can prevent one feature representation overshadowing the other due to their different value ranges or feature dimensions. The term $\|\boldsymbol{s}_j\|_2^2$ and $\|\boldsymbol{s}_j\|_1$ are the ℓ_2-norm and ℓ_1-norm regularization terms, respectively, and β and γ are their regularization parameters. A larger regularization parameter imposes a severe regularization. The ℓ_1-norm is introduced to get a sparse solution of $\boldsymbol{S}$ [22], which can make the updating process of Equation (2) very fast, especially when dealing with a large number of instances. It also has effect on noise removal, which has been extensively used in image processing [23]. The ℓ_2-norm can restrict parameter range and prevent the model from overfitting. The two regularization terms together lead the optimization problem to the elastic net [24], [25], which balances between the lasso using the ℓ_1-norm and ridge regression using the ℓ_2-norm. The constraint $diag(\boldsymbol{S}) = 0$ is applied to avoid trivial solutions [20], that is the optimal $\boldsymbol{S}$ is an identical matrix such that an instance is always best related to itself and not related to any other instance. The constraint $(\alpha_j^t)^2 + (\alpha_j^v)^2 = 1$ is to balance the weight between the two modalities.

$$\begin{aligned} \min_{\boldsymbol{s}_j, \alpha_j^t, \alpha_j^v} \quad & \frac{(\alpha_j^t)^2}{2} \|\boldsymbol{x}_j^t - \boldsymbol{X}^t \boldsymbol{s}_j\|_2^2 + \frac{(\alpha_j^v)^2}{2} \|\boldsymbol{x}_j^v - \boldsymbol{X}^v \boldsymbol{s}_j\|_2^2 \\ & + \frac{\beta}{2} \|\boldsymbol{s}_j\|_2^2 + \gamma \|\boldsymbol{s}_j\|_1 \\ s.t. \quad & s_{jj} = 0 \\ & (\alpha_j^t)^2 + (\alpha_j^v)^2 = 1 \end{aligned} \tag{3}$$

Using Lagrange multiplier, solving Equation (3) is equivalent to solve Equation (4). For simplicity, $\boldsymbol{x}_j = [(\alpha_j^t \boldsymbol{x}_j^t)^T, (\alpha_j^v \boldsymbol{x}_j^v)^T]^T$ is used in the following derivation whenever applicable, where $\boldsymbol{X} \in \mathbb{R}^{M \times N}$ and $M = M^t + M^v$. Let J denote the cost function in Equation (4), which is depended on $\boldsymbol{s}_j$, α_j^t and α_j^v. All the terms in J are differentiable except $\|\boldsymbol{s}_j\|_1$. A global minimum of J can be found using coordinate descent [26]. The partial derivative of J with respect to the i-th entry of $\boldsymbol{s}_j$ is derived as Equation (5), and the update form s_{ij} is shown in Equation (6), where Υ is the soft-thresholding operator. Similarly, taking the partial derivative of J with respect to α_j^t and α_j^v can get the update form of these two variables, as shown in Equation (7). Repeat the update of each of the variables for a certain number of times, together with the partial derivative of J with respect to λ or equivalently the constraint $(\alpha_j^t)^2 + (\alpha_j^v)^2 = 1$, J is converged and the optimal values of $\boldsymbol{s}_j$, α_j^t and α_j^v are reached. This update process is terminated if one of the two criteria is met: 1) the maximum number of iterations; 2) the acceptable tolerance between two iterations.

$$\begin{aligned} \min_{\boldsymbol{s}_j, \alpha_j^t, \alpha_j^v} \quad & \frac{1}{2} \|\boldsymbol{x}_j - \boldsymbol{X}\boldsymbol{s}_j\|_2^2 + \frac{\beta}{2} \|\boldsymbol{s}_j\|_2^2 + \gamma \|\boldsymbol{s}_j\|_1 \\ & + \lambda((\alpha_j^t)^2 + (\alpha_j^v)^2 - 1) \end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\partial J}{\partial s_{ij}} &= -\sum_{h=1}^{h=M} x_{hi}(x_{hj} - \sum_{g=1}^{g=N} x_{hg}s_{gj}) + \beta s_{ij} + \gamma \\ &= -\sum_{h=1}^{h=M} x_{hi}(x_{hj} - \sum_{g\neq i} x_{hg}s_{gj} - x_{hi}s_{ij}) + \beta s_{ij} + \gamma \\ &= -\sum_{h=1}^{h=M} x_{hi}(x_{hj} - \sum_{g\neq i} x_{hg}s_{gj}) + \sum_{h=1}^{h=M} x_{hi}^2 s_{ij} \\ &\quad + \beta s_{ij} + \gamma \\ &= -\sum_{h=1}^{h=M} x_{hi}(x_{hj} - \sum_{g\neq i} x_{hg}s_{gj}) \\ &\quad + (\sum_{h=1}^{h=M} x_{hi}^2 + \beta)s_{ij} + \gamma \end{aligned} \tag{5}$$

$$s_{ij} \leftarrow \frac{\Upsilon(\sum_{h=1}^{h=M} x_{hi}(x_{hj} - \sum_{g\neq i} x_{hg}s_{gj}), \gamma)}{\sum_{h=1}^{h=M} x_{hi}^2 + \beta}, \quad where$$

$$\Upsilon(z,\gamma) = \begin{cases} z-\gamma & \text{if } z>0 \text{ and } |z|>\gamma \\ z+\gamma & \text{if } z<0 \text{ and } |z|>\gamma \\ 0 & \text{if } |z|\leq\gamma \end{cases} \tag{6}$$

$$\alpha_j^t \leftarrow \frac{-2\lambda}{\|\boldsymbol{x}_j^t - \boldsymbol{X}^t\boldsymbol{s}_j\|_2^2}$$

$$\alpha_j^v \leftarrow \frac{-2\lambda}{\|\boldsymbol{x}_j^v - \boldsymbol{X}^v\boldsymbol{s}_j\|_2^2} \tag{7}$$

Borrowing the concept of the reconstruction error [27] from the sparse representation, an instance can be reconstructed from other instances weighted by the coefficients between them. Comparing to the sparse representation that introduces the ℓ_1-norm to get a sparse solution of $\tilde{\boldsymbol{s}}$, SLI adds the ℓ_2-norm to further prevent the model from overfitting. A classifier similar to the sparse representation-based classifier can be built using the reconstruction error generated from the instances of different classes. Given a test instance $\boldsymbol{y}$, Equation (8) measures the test error for class c, where $\tilde{\boldsymbol{s}}$ is the coefficient between $\boldsymbol{y}$ and the positive instances $\boldsymbol{X_c}$ of class c. The probability of $\boldsymbol{y}$ belonging to class c, denoted as $prob_c(y)$, is inversely proportional to $err_c(y)$, which can be calculated by various mapping functions such as Gaussian kernel.

$$err_c(y) = \|\boldsymbol{y} - \boldsymbol{X_c}\tilde{\boldsymbol{s}}\|_2^2. \tag{8}$$

IV. EXPERIMENT

For image semantic concept retrieval, two benchmark datasets MIRFLICKR-25000 collection [28] and NUS-WIDE-LITE [29] are widely used. Section 4.1 and Section 4.2 present the detailed evaluation results. Mean average precision (MAP) is one of the most widely used metric in information retrieval. It is the mean of the average precision scores for each query, while average precision (AP) is computed as the average value of the precision-recall curve, as shown in Equation (9). Q is the total number of queries, and TP@n is the number of true positive at cut-off n. P@i is the precision at cut-off i in the ranking list and $\Delta(i)$ is an indicator function equaling 1 if the item at rank i is a relevant one, 0 otherwise. n can be set to such as 5, 10 and 100 depending on the circumstances. If all the retrieval results are considered, then AP@all and MAP@all can be used.

$$MAP@n = \frac{\sum_{q=1}^{q=Q} AP@n}{Q}$$

$$where \quad AP@n = \frac{\sum_{i=1}^{i=n} P@i \times \Delta(i)}{TP@n}. \tag{9}$$

A. MIRFLICKR-25000

MIRFLICKR-25000 collection contains 25000 images and their associated tags from the Flickr website. 38 concepts are manually annotated for research purposes. Their concept IDs and names are listed in Table 1. It includes two types of labels: potential labels (24 concepts out of 38) and relevant labels (14 concepts out of 38). Potential labels of a concept are given to images as long as the concept is visible or applicable to some extent, while relevant labels are given to images only if the annotator found the image really relevant to his/her interpretation regarding to a certain concept. For completeness, all 38 concepts are used in the experiment. A standard way to split the training and test sets are defined by this collection, 15000 out of 25000 are the training data and the rest 1000 are the test data. Their positive to negative (PN) ratios of the 38 concepts are depicted in Figure 4. The concept name ends with "_r1" denotes the concept having relevant labels. As can be seen, the data in MIRFLICKR-25000 ranges from highly imbalanced ones (low PN ratios) to relatively balanced ones (high PN ratios).

To build the content modality, 4 types of features are extracted from the 25000 images. They are color moment in the YCbCr space [30], Local Binary Patterns (LBP) [31],

TABLE 1. Names of the 38 concepts from MIRFLICKR-25000.

1	animals	11	dog_r1	21	night	31	sea_r1
2	baby	12	female	22	night_r1	32	sky
3	baby_r1	13	female_r1	23	people	33	structures
4	bird	14	flower	24	people_r1	34	sunset
5	bird_r1	15	flower_r1	25	plant_life	35	transport
6	car	16	food	26	portrait	36	tree
7	car_r1	17	indoor	27	portrait_r1	37	tree_r1
8	clouds	18	lake	28	river	38	water
9	clouds_r1	19	male	29	river_r1		
10	dog	20	male_r1	30	sea		

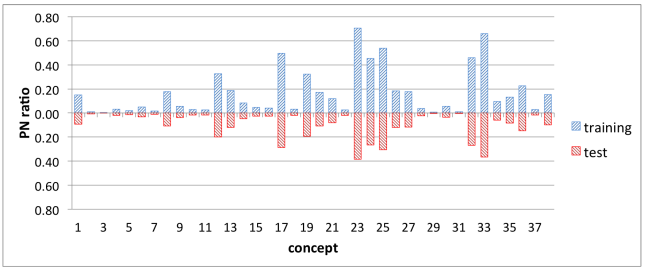

FIGURE 4. The PN ratios of the 38 concepts from MIRFLICKR-25000.

histogram of oriented gradients (HOG) [32], and haar wavelets [33]. The total number of visual feature dimensions is 551. For each image, tags are assigned by Flclikr users, which probably contain typos, non-English words, unrelated tags, etc. Standard procedures such as stop word removal and stemming are applied to these tags. English word validation is also used to validate each word by checking whether it exists in Wordnet [34], which is a large lexical database of English. Textual features are extracted from 10055 unique terms after this preprocessing. To maintain the textual features in the same scale as the visual features, the top 500 terms with the highest χ^2 values are selected. The binary representation is used to indicate the presence or absence of a term.

As one of the most popular classifiers, the logic regression (LR) model is adopted as the basic classifier to evaluate the content and the context modality separately, denoted as "content" and "context". To show that content and context modalities can often provide complementary information, the results of AP@10 using content and context feature representations alone are displayed in Figure 5. The results of AP@10 are shown because this complementary characteristic is more notable in high ranked instances. As can also be observed from this figure, the context modality performs much better on some concepts than the content modality, such as "baby", "bird", "car", 'and 'dog". However, on concept "river_r1", the context modality completely fails. Therefore, our motivation of integrating content and context modalities can be proved on this dataset.

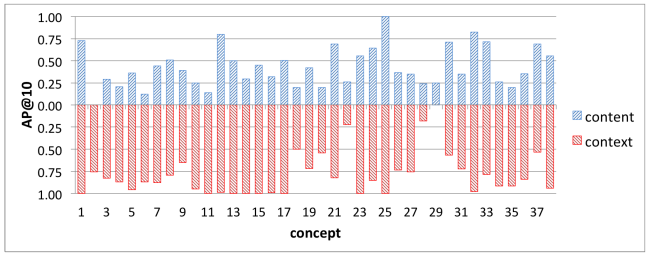

FIGURE 5. AP@10 of content and context modalities on the 38 concepts.

The parameters that need to be tuned in the SLI model are β and λ as shown in Equation (3), which are the parameters of the ℓ_2-norm and the ℓ_1-norm respectively. The grid search approach is adopted, where the search range for β is from 0.001 to 10.0 with points at $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10\}$, and the search range for λ is from 0.001 to 0.5 with points at $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. Using three-fold cross-validation on the training dataset achieves relatively good results when $\beta = 1.0$ and $\lambda = 0.01$. Meanwhile, SLI does not depend on the initial state of $\boldsymbol{S}$. In our implementation, s_{ij} is initialized with normal distributed (Gaussian) noise of 0 mean and 0.1 standard deviation. However, the constraint $diag(\boldsymbol{S}) = 0$ is still imposed in the initialization. For the stopping criteria, the maximum number of iterations is set to 10000 and the acceptable tolerance between two iterations is set to 0.0001. Figure 6 shows the MAP of the 38 concepts. EF_LR represents the early fusion approach, which directly concatenates the content and context feature representations, and applies LR to the concatenated feature representation. LF_LR represents the late fusion approach, which first adopts logistic regression for local decisions, and then applies it again to fuse local decisions from the two modalities. SLI denotes the proposed sparse linear integration model. As can be seen from the figure, the results of "content" have relatively low precision values in the top retrieved images, but the precision values increase as more images are included till the top 100 retrieved instances. The results of "context" show an opposite behavior, which achieves high precision values in the top retrieved images and starts to decrease when more images are included. It can be seen that EF_LR achieves stable results than methods rely on a single modality. However, on MAP@10 "context" performs slightly better than EF_LR, and on MAP@100 and MAP@all, "content" gives a much better performance. Look more closely at the figure, we can see that EF_LR actually produces an "averaged" results between "content" and "context". We also notice that LF_LR achieves better results compared to EF_LR, and it does not suffer from the "averaged" problem. On the other hand, SLI achieves the best performance on all metrics, and the relevant improvements on MAP@10, MAP@20, MAP@50, MAP@100 and MAP@all compared to EF_LR are 17.8%, 13.0%, 11.2%, 10.8%, and 8.9%, respectively. The corresponding improvements compared to "LF_LR" are 9.3%, 6.1%, 7.6%, 3.9%, and 4.5%.

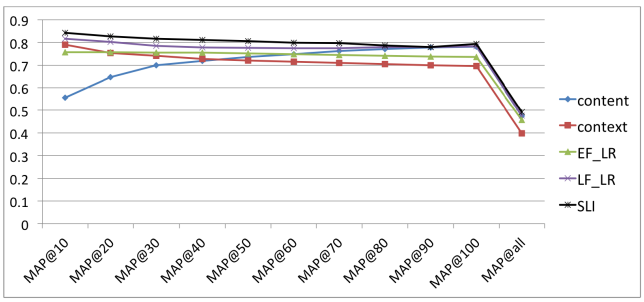

FIGURE 6. Comparison results of MAP on MIRFLICKR-25000.

A similar work [13] is discussed in Section 2. It uses matrix factorization (MF) to integrate the content and context feature representations. Compared to the matrix factorization technique, the SLI model does not need to decide the latent factor, which could cause information loss due to the low-rank approximation. In addition, the sparsity also keeps a low computation complexity. Experiments conducted in [13] used the same training and test datasets but a different set of features. For the content modality, a dictionary of 2000 visual features is used while for the context modality, 1391 keywords are used by keeping those keywords appeared more than 20 times, and idf weights are used instead of binary values. Due to the difficulty of reproducing the exact feature sets and the fact that the performance also depends on

the classifier it adopts, it is impractical to evaluate their fusion model alone in the same setting as in [13]. As an alternative, we compare the performance of the whole process. Given a much larger feature space, their reported results as shown in Table 2 are quite low. MF-visual denotes the backprojected content feature representation, MF-textual denotes the backprojected context feature representation, and MF-latent is the weighted concatenation of the two backprojected representations. The performance of two other methods in [13] are also reported, The visual search method uses the original visual representation, and the semantic embedding method finds a semantic transformation from textual features to visual features. The metrics adopted in their experiments are MAP@all and P@10. P@10 is simply the precision value of the first 10 results. Thus, we also calculate the P@10 values of "content", "context", EF_LR, LF_LR, and SLI, which are 0.392, 0.695 , 0.724, 0.807 and 0.835. The P@10 values of "content" are smaller than those of "MF-visual" and "visual search", which are all generated from the content modality only. This is probably due to the common visual features we used, but our visual feature dimension is only about a quarter of theirs. The MAP@all value of "content" outperforms all their methods. On the context side, the 500 terms we extracted are much more effective compared to the 1391 terms used in their experiments. "context" alone outperforms their methods in terms of both P@10 and MAP@all.

TABLE 2. MAP of the 38 concepts reported in work [13].

	P@10	MAP@all
MF-visual	0.426	0.315
MF-textual	0.374	0.287
MF-latent	0.440	0.288
visual search	0.526	0.309
semantic embedding	0.258	0.227

B. NUS-WIDE-LITE

The same evaluation is also conducted on NUS-WIDE-LITE, which contains 55, 615 images with associated tags from the Flickr website. This dataset has also been divided by the dataset provider into training and test sets in advance, where 27, 807 images are used as the training data and the test data is composed of the remaining 27, 808 images. Some low-level features are provided, including color histogram, wavelet texture, and etc. The low-level features used here in this experiment are 64-dimensional color histogram in LAB color space and 128-dimensional wavelet texture, which are basic features that are commonly extracted to analyze the content of images. 81 concepts provided by this dataset are listed in Table 3. It also provides 1, 000 frequent tags that are used as the context modality, but they contain much less noisy tags compared to MIRFLICKR-25000. The PN ratios of the concepts are shown in Figure 7, It can be seen from this figure that most of the concepts are very imbalanced in that the number of positive images (images containing a target concept) divided by the number of negative images (images without a target concept) is smaller than 0.05.

TABLE 3. Names of the 81 concepts from NUS-WIDE-LITE.

1	airport	28	frost	55	sign
2	animal	29	garden	56	sky
3	beach	30	glacier	57	snow
4	bear	31	grass	58	soccer
5	birds	32	harbor	59	sports
6	boats	33	horses	60	statue
7	book	34	house	61	street
8	bridge	35	lake	62	sun
9	buildings	36	leaf	63	sunset
10	cars	37	map	64	surf
11	castle	38	military	65	swimmers
12	cat	39	moon	66	tattoo
13	cityscape	40	mountain	67	temple
14	clouds	41	nighttime	68	tiger
15	computer	42	ocean	69	tower
16	coral	43	person	70	town
17	cow	44	plane	71	toy
18	dancing	45	plants	72	train
19	dog	46	police	73	tree
20	earthquake	47	protest	74	valley
21	elk	48	railroad	75	vehicle
22	fire	49	rainbow	76	water
23	fish	50	reflection	77	waterfall
24	flags	51	road	78	wedding
25	flowers	52	rocks	79	whales
26	food	53	running	80	window
27	fox	54	sand	81	zebra

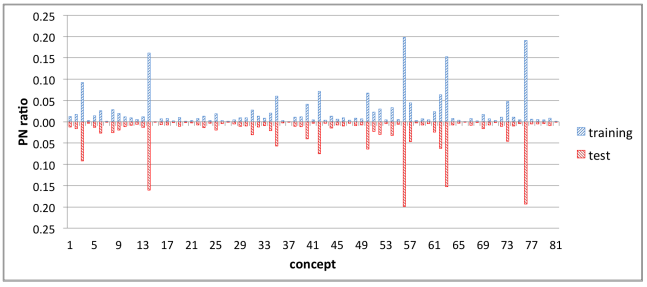

FIGURE 7. The PN ratio of all 81 concepts from NUS-WIDE-LITE.

Retrieving positive instances from highly imbalanced dataset is a very challenging issue in the area of multimedia semantic information retrieval.

The results of AP@10 using logistic regression based on the content and context feature representations are displayed in Figure 8. As can be seen, for the NUS-WIDE-LITE dataset, the context modality generates much better performance, which is due to the fact that the tags provided by this dataset are already been cleaned and thus the quality is high. On the other hand, the performance generated from the content feature representation is considerably inferior compared to that of the context feature representation. However, given the high quality of the context modality, there still exist concepts that visual features are very useful, such as concept No.41 "nighttime". This finding is also intuitive since color-based visual features are expected to play an important role in discriminating this concept. From NUS-WIDE-LITE, the same conclusion is drawn that the retrieval performance can be greatly enhanced if the two modalities are properly integrated.

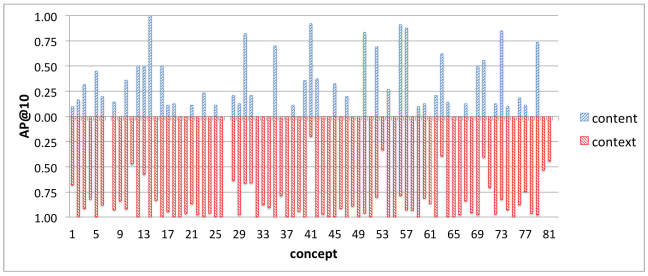

FIGURE 8. AP@10 of content and context modalities on the 81 concepts.

Similar to Figure 6, Figure 9 shows the comparison results of the MAP values. For SLI, $\beta = 1.0$ and $\lambda = 0.1$ are found using the same parameter tuning approach, the rest of the parameter settings are the same as they are for MIRFLICKR-25000 dataset. As can be seen, "context" performs much better than "content". This is because the tags provided by this dataset has been cleaned and thus contains much less noise. The low performance of "content" also causes the MAP values of the EF_LR smaller than those of "context". It can be observed that EF_LR is subject to the modality having the worst performance. However, it is able to achieve slightly better results than "context" at MAP@100 and MAP@all. The performance of LF_LR is much robust comparing to that of EF_LR. It constantly generates better MAP values than those rely on a single modality. SLI still achieves the best results on this dataset, and the relative improvements on MAP@10, MAP@20, MAP@50, MAP@100 and MAP@all comparing to EF_LR are 19.8%, 20.4%, 12.7%, 9.7% and 11.5%. The corresponding improvements compared to LF_LR are 4.5%, 7.1%, 5.3%, 3.6% and 5.1%, which show a stable improvement of around 4%.

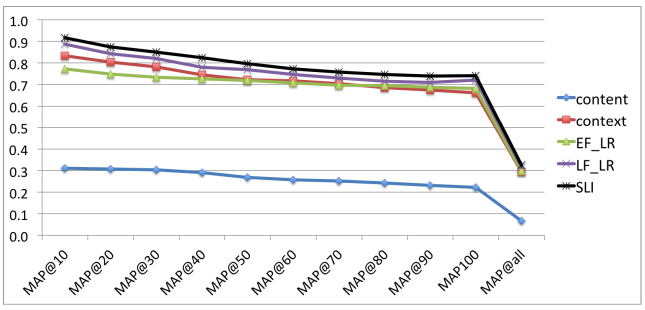

FIGURE 9. Comparison results of MAP on NUS-WIDE-LITE.

SLI is also evaluated against several popular fusion approaches, including methods using "minimum" (min), "maximum" (max), "median", and "average" rules. Here, majority voting is not included since it requires hard decision on class labels. The super kernel fusion (SKF) method [35] is also compared, as well as one of our previously work [36], which considers adjustment, reliability, and correlation of the intervals to the target concept (denoted as ARC). The local decisions of these methods are obtained using SVM [37]. The experiment setup is based on the experiment

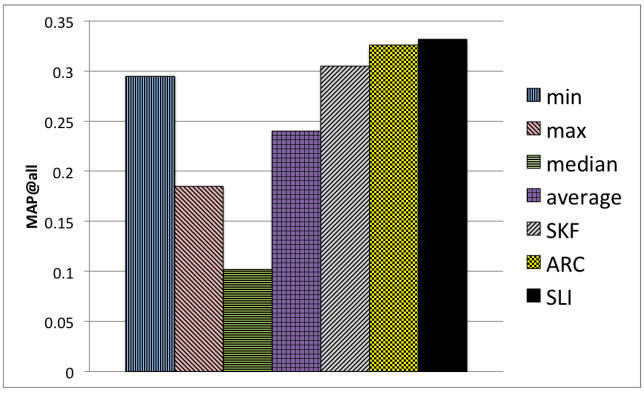

FIGURE 10. Comparison results of MAP on NUS-WIDE-LITE.

conducted in [36]. The comparison MAP@all values on the NUS-WIDE-LITE dataset are shown in Figure 10. As can be seen, median fusion gives the worst performance and is outperformed by SLI with a large margin of 23%. ARC produces the second best result but is still 1.0% lower than SLI. The min fusion method shows fairly good results and is better than the average and max fusion methods.

V. CONCLUSION

In this paper, a sparse linear integration model called SLI is proposed to integrate the content and the context modalities for semantic concept retrieval. The integration process is formulated into an optimization problem that aims to approximate an instance using a sparse linear combination of other instances, which can be solved by coordinate descent and soft-thresholding. Classification is performed by using the positive training instances of a class to reconstruct a test instance, and the smaller the reconstruction error is, the more likely that the test instance belongs to this class. Evaluation of the SLI model is conducted on two benchmark image datasets. Comparison methods include a logistic regression on each of the modality alone as well as the early fusion and late fusion approaches. Results from recent publications using the same datasets are also included in comparison. Approaches based on two modalities demonstrate superiority over a single modality based methods in general. SLI shows promising results by outperforming the early fusion and late fusion approaches based on logistic regression.

As a current limitation, SLI requires the feature representation from each modality to be in the same scale. That is, their feature dimensions should be similar; otherwise, one feature representation would overshadow the other. This would cause the model to lean toward the one with a higher dimension, which means the one with higher dimensions would contribute more to the learned coefficients. Therefore, certain techniques such as feature selection need to be applied first in order to make the dimensions of different feature representations be in the same scale.

REFERENCES

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, Feb. 2006.

[2] A. Yoshitaka and T. Ichikawa, "A survey on content-based retrieval for multimedia databases," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 1, pp. 81–93, Jan./Feb. 1999.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[4] H. Li, X. Wu, Z. Li, and W. Ding, "Group feature selection with streaming features," in *Proc. Int. Conf. Data Mining*, Dec. 2013, pp. 1109–1114.

[5] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang, "Event-driven semantic concept discovery by exploiting weakly tagged internet images," in *Proc. Int. Conf. Multimedia Retr.*, 2014, pp. 1:1–1:8.

[6] Q. Zhu, Z. Li, H. Wang, Y. Yang, and M.-L. Shyu, "Multimodal sparse linear integration for content-based item recommendation," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2013, pp. 187–194.

[7] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.

[8] T. Meng and M.-L. Shyu, "Model-driven collaboration and information integration for enhancing video semantic concept detection," in *Proc. IEEE 13th Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2012, pp. 144–151.

[9] D. Liu and M.-L. Shyu, "Effective moving object detection and retrieval via integrating spatial-temporal multimedia information," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2012, pp. 364–371.

[10] M. Žitnik and B. Zupan. (2013). "Data fusion by matrix factorization." [Online]. Available: http://arxiv.org/abs/1307.0803

[11] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 399–402.

[12] K. Nagel, S. Nowak, U. Kühhirt, and K. Wolter, "The Fraunhofer IDMT at ImageCLEF 2011 photo annotation task," in *Proc. CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[13] J. C. Caicedo and F. A. González, "Multimodal fusion for image retrieval using matrix factorization," in *Proc. 2nd ACM Int. Conf. Multimedia Retr.*, 2012, pp. 56:1–56:8.

[14] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," in *Proc. 16th Comput. Vis. Winter Workshop*, 2011, pp. 1–8.

[15] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui, "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*, vol. 76, no. 1, pp. 50–60, Jan. 2012.

[16] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2211–2268, Feb. 2011.

[17] S. Yu *et al.*, "L_2-norm multiple kernel learning and its application to biomedical data fusion," *BMC Bioinform.*, vol. 11, no. 1, p. 309, 2010.

[18] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C. F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 563–574, Jun. 2012.

[19] X. Ning and G. Karypis, "SLIM: Sparse linear methods for top-N recommender systems," in *Proc. IEEE 11th Int. Conf. Data Mining (ICDM)*, Dec. 2011, pp. 497–506.

[20] X. Ning and G. Karypis, "Sparse linear methods with side information for top-N recommendations," in *Proc. 6th ACM Conf. Recommender Syst.*, 2012, pp. 155–162.

[21] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[23] Y. Traonmilin, S. Ladjal, and A. Almansa, "Outlier removal power of the L1-norm super-resolution," in *Scale Space and Variational Methods in Computer Vision* (Lecture Notes in Computer Science), vol. 7893, A. Kuijper, K. Bredies, T. Pock, and H. Bischof, Eds. Berlin, Germany: Springer-Verlag, 2013, pp. 198–209.

[24] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., Ser. B*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[25] A. Huang, S. Xu, and X. Cai, "Empirical Bayesian elastic net for multiple quantitative trait locus mapping," *Heredity*, vol. 114, no. 1, pp. 107–115, 2014. [Online]. Available: http://www.nature.com/hdy/journal/vaop/ncurrent/full/hdy201479a.html

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.

[27] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[28] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Inf. Retr. (MIR)*, 2008, pp. 39–43.

[29] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, Jul. 2009, pp. 1–9.

[30] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the HSV color space for image retrieval," in *Proc. Int. Conf. Image Process. (ICIP)*, 2002, pp. 589–592.

[31] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[33] D. N. Verma and V. Maru, "An efficient approach for color image retrieval using Haar wavelet," in *Proc. Int. Conf. Methods Models Comput. Sci.*, Dec. 2009, pp. 1–5.

[34] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.

[35] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proc. 12th Annu. ACM Int. Conf. Multimedia (MM)*, 2004, pp. 572–579.

[36] C. Chen, Q. Zhu, L. Lin, and M.-L. Shyu, "Web media semantic concept retrieval via tag removal and model fusion," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, pp. 61:1–61:22, Sep. 2013.

[37] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

QIUSHA ZHU has been a Data Scientist in graph database for music recommendation with Senzari Inc., Miami, FL, USA, since 2014. She was a Research Intern in algorithms for movie recommendation with TCL Research America, San Jose, CA, USA, from 2012 to 2013. She received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, USA, in 2014, and the M.S. degree in electronics engineering from the School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2009. Her research interests include multimedia information retrieval, information fusion, and recommender systems, in particular, how to leverage distributed and parallel computing to handle big data.

MEI-LING SHYU has been a Full Professor with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, USA, since 2013, where she has been an Associate/Assistant Professor of Electrical and Computer Engineering since 2000. She received the Ph.D. degree from the School of Electrical and Computer Engineering and three master's degrees from Purdue University, West Lafayette, IN, USA. Her research interests include multimedia data mining, management and retrieval, and security. She has authored or co-authored two books and over 230 technical papers. She was a recipient of the Computer Society Technical Achievement Award in 2012, the ACM Distinguished Scientists Award in 2012, the Best Paper Awards from the IEEE International Symposium on Multimedia in 2013 and the IEEE International Conference on Information Reuse and Integration in 2014 and 2012, the Best Published Journal Article in the *International Journal of Multimedia Data Engineering and Management* Award in 2010, and the Best Student Paper Award with her student from the Third IEEE International Conference on Semantic Computing in 2009. She is a fellow of the Society of Information Reuse and Integration.