Enhancing Micro-video Understanding by Harnessing External Sounds

Liqiang Nie
ShanDong University
nieliqiang@gmail.com

Xiang Wang
National University of Singapore
xiangwang@u.nus.edu

Jianglong Zhang
Communication University of China
zhangjianglong135@126.com

Xiangnan He
National University of Singapore
xiangnanhe@gmail.com

Hanwang Zhang
Columbia University
hanwangzhang@gmail.com

Richang Hong
Hefei University of Technology
hongrc.hfut@gmail.com

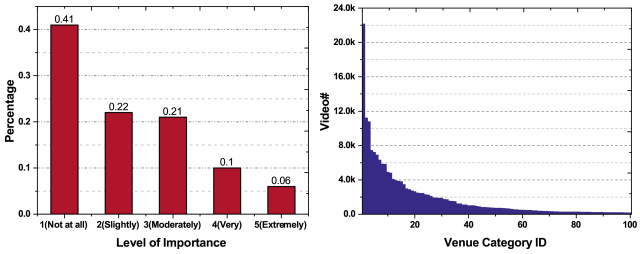
Qi Tian
University of Texas at San Antonio
qi.tian@utsa.edu

ABSTRACT

Different from traditional long videos, micro-videos are much shorter and usually recorded at a specific place with mobile devices. To better understand the semantics of a micro-video and facilitate downstream applications, it is crucial to estimate the venue where the micro-video is recorded, for example, in a concert or on a beach. However, according to our statistics over two million micro-videos, only 1.22% of them were labeled with location information. For the remaining large number of micro-videos without location information, we have to rely on their content to estimate their venue categories. This is a highly challenging task, as micro-videos are naturally multi-modal (with textual, visual and, acoustic content), and more importantly, the quality of each modality varies greatly for different micro-videos.

In this work, we focus on enhancing the acoustic modality for the venue category estimation task. This is motivated by our finding that although the acoustic signal can well complement the visual and textual signal in reflecting a micro-video's venue, its quality is usually relatively lower. As such, simply integrating acoustic features with visual and textual features only leads to suboptimal results, or even adversely degrades the overall performance (*cf.* the barrel theory). To address this, we propose to compensate the shortest board — the acoustic modality — via harnessing the external sound knowledge. We develop a deep transfer model which can jointly enhance the concept-level representation of micro-videos and the venue category prediction. To alleviate the sparsity problem of unpopular categories, we further regularize the representation learning of micro-videos of the same venue category. Through extensive experiments on a real-world dataset, we show that our model significantly outperforms the state-of-the-art method [47] in terms of both Micro-F1 and Macro-F1 scores by leveraging the external acoustic knowledge.

MM '17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123313



(a) Distribution of User Ratings (b) Disbtribution of Venue Categories

Figure 1: (a) User study of importance of acoustic signal. (b) Distribution of top 100 venue categories of the micro-video benchmark data [47].

CCS CONCEPTS

•Information systems → Multimedia information systems;

KEYWORDS

External Sound Knowledge; Micro-video Categorization; Deep Neural Network; Representation Learning;

1 INTRODUCTION

The emerging micro-video sharing platforms, such as Vine[1], Snapchat[2], and Instagram[3], enable users to shoot, capture, and share micro-videos of their daily life at any time and any place. Different from traditional long videos, micro-videos, lasting for 6-15 seconds, greatly cater to users' narrow attention spans, and they are usually recorded at a specific place with smart mobile devices. The success of these platforms is taking the media world by storm and benefits many potential applications, such as marketing and advertising [7]. However, as a new media form, research studies on micro-video understanding, such as venue estimation, event detection, and object tracking, are relatively sparse.

Recently, Zhang *et al.* [47] conducted a preliminary study on micro-video understanding by estimating their venue categories. It utilizes the consensus information on visual, acoustic, and textual

[1]https://vine.co.
[2]https://www.snapchat.com.
[3]https://www.instagram.com.

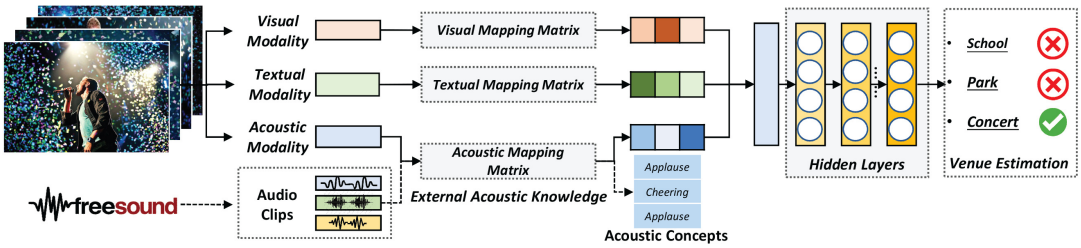


Figure 2: Schematic illustration of our proposed deep transfer model. It transfers knowledge from external sound clips to strengthen the description of the internal acoustic modality in micro-videos. Meanwhile, it conducts a deep multi-modal fusion towards venue category estimation.

modalities of micro-videos to recognize the venue information. The multi-modal method, unfortunately, overlooks the fact that the quality of each modality varies dramatically. According to their experiments [47], the acoustic modality demonstrates the weakest capability in indicating the venue information. Ignoring the varying quality of different modalities may cause the negative cask effect, resulting in suboptimal prediction performance.

To gain deep insights into micro-videos, we performed a user study to explore the influence of acoustic signal on estimating the venue category. Given 100 micro-videos randomly sampled from Vine, five volunteers were invited to guess the venue category of each micro-video by listening to its sound only, without knowing its textual and visual content. Subsequently, they were asked to rate the importance of sound of each micro-video with a score 1 to 5, where a higher score indicates that the acoustic signal is more informative to reflect the venue category. As shown in Figure 1(a), we have two key observations: 1) for 59% micro-videos, the acoustic modality can benefit the venue category estimation to a certain extent. This points to the positive effect of acoustic concepts; for example, recognizing *bird chirps* or *crowds cheering* from sound is helpful for estimating *a park* or *concert*. However, 2) the acoustic signal of 84% micro-videos are insufficient to reflect the venue category accurately (scores below 4), pertaining to the inherent noise and low quality. This study lends support to the usefulness of acoustic information of micro-videos, however, they need to be further refined for enhancing performance.

Leveraging the rich, external sound knowledge to compensate the internal acoustic signal is an intuitive thought. Nevertheless, it is non-trivial to implement due to the following challenges:

- Since most micro-videos record events of users' daily life, we need to learn high-level acoustic concepts to better distinguish events [35]. However, to our knowledge, there is no suitable sound data for micro-videos, as existing labelled data are either too small to cover the common acoustic concepts [30] or constructed from videos of limited event categories [3, 33].
- External sounds are unimodal data; whereas micro-videos unify textual, visual, and acoustic modalities to describe a real-life event. It is technically challenging to fuse the unimodal sound data to improve the learning of multi-modal video data.
- According to our statistics, we observe a severe sparsity problem of unpopular categories (*cf.* Figure 1(b)). The insufficient training samples easily result in a poor classifier, which tends to classify an unseen micro-video into the dominated categories.

To address these challenges, we first construct 313 high-level acoustic concepts that cover most common real-life sounds; we then collect 43,868 sound clips from Freesound[4] based on the acoustic concepts. We design a *Deep trAnsfeR modEl* (DARE), which jointly leverages external sounds to strengthen the acoustic concept learning and the category similarity to alleviate the sparsity problem. Figure 2 illustrates the workflow of our DARE approach. Specifically, we first extract features for each modality, and then project the features of each modality with a dedicated mapping matrix to obtain the high-level representations. To transfer the external sound knowledge, we apply the same acoustic feature extractor on the labelled audio clips and use the same mapping matrix as the acoustic modality. Following that, we concatenate the representations of three modalities and feed it into a deep neural network with multiple hidden layers, which can capture the non-linear correlations among concepts. To alleviate the sparsity problem of unpopular categories, we preserve the similarity of micro-videos according to their venue categories. Formally, we encourage the micro-videos within the same category to have similar representations in the latent space; meanwhile, the ones from different categories are enforced to be dissimilar with each other. As such, the representation learning of unpopular categories can considerably benefit from that of popular categories and thus boost the representation learning. We ultimately feed the fused representations into a prediction function to estimate the venue categories. We validate our DARE model over a publicly accessible benchmark dataset. Extensive experiments demonstrate that our model can yield promising performance.

The main contributions of this work are threefold:

- We construct a set of acoustic concepts with corresponding sound clips, covering most of the frequent real-life sounds. We have released this dataset and the source codes of this work to facilitate the research community[5].
- We build a deep transfer model to estimate the venue categories of micro-videos. It is capable of seamlessly transferring the external sound knowledge to enhance the acoustic modality description.
- We alleviate the sparsity problem of unpopular categories by regularizing the similarity among categories, and then obtain the discriminative and conceptual representation of micro-videos by modality-aware mapping functions.

[4]https://freesound.org/.
[5]https://goo.gl/DCtVE6.

The rest of this paper is structured as follows. In Section 2, we review the related work. Section 3 and 4 respectively detail our data collection and our proposed DARE model. We conduct experiments and analyze the results in Section 5, followed by conclusion and future work in Section 6.

2 RELATED WORK

Our work is closely related to multimedia location estimation, dictionary learning, and acoustic concept detection.

2.1 Multimedia Location Estimation

Roughly speaking, pioneer efforts on multimedia location estimation can be grouped into two categories: unimodal venue estimation [4, 8, 15] and multi-modal venue estimation [9, 13, 47]. Approaches in the former category extract a rich set of visual features from images and leverage the visual features to train either shallow or deep models to estimate the venues of the given images. Beyond the unimodal venue estimation which only takes the visual information into account, multi-modal venue estimation works infer the geo-coordinates of the video recording places by fusing the textual and visual/acoustic cues. The principle is that integration of multiple modalities can lead to better results, and it is consistent with the old saying *two heads are better than one* cues [20, 21, 37, 38]. However, multi-modal venue estimation is still at its infant stage, and few of them pay attention to the cask effect phenomenon, let alone borrowing knowledge from external sources.

2.2 Dictionary Learning

We claim that our DARE model is related to dictionary learning, since they both learn conceptual representations. Dictionary learning aims to find a dictionary of atoms (concepts), in which each sample admits a sparse representation in the form of a linear combination of atoms. Existing efforts are in either unsupervised or supervised settings. The unsupervised one aims to reconstruct the original signals as precise as possible via minimizing the reconstruction error. They achieved promising performance in the reconstruction tasks, such as denoising [11], restoring [27], and coding [25]. Despite their value in the reconstruction tasks, they are unfavorable in classification tasks [42]. This motivates the development of supervised dictionary learning [28, 48], which leverages the class labels in the training set to build a more discriminative dictionary for the particular classification task at hand. They have been well adapted to many applications with better performance, such as painting style inferring [22] and image classification [24]. Different from prior efforts that have sparse constraints and have to learn dictionaries, our method uses mapping functions to project the low-level features to high-level conceptual representations.

2.3 Acoustic Concept Detection

Acoustic concept detection on the user-generated videos is a relatively new field in multimedia community [34], composing of the data-driven [5, 6] and task-driven [1, 33] approaches from the perspective of modeling acoustic concepts. The main motivation of acoustic concept detection is that audio analysis provides a complementary information to detect the specific events that are hardly identified with visual cues. Recent studies [41] have shown that detecting sound events to bridge the gap between the low-level features and the high-level semantics outperforms the pure feature-based approaches. Different from acoustic concept detection, we target at constructing a knowledge base of acoustic concepts and leveraging such base to strengthen the representation learning of micro-videos.

3 DATA COLLECTION

In this section, we describe the details of the datasets of the micro-videos and our constructed external sounds.

3.1 Micro-video Dataset

To validate our work, we leveraged a public benchmark micro-video dataset[6]. Micro-videos in this dataset were collected from Vine and exclusively distributed in 442 venue categories. We filtered out those categories with less than 50 micro-videos following [47]. We ultimately left 270, 145 micro-videos over 188 venue categories. Each micro-video is described by a rich set of features, namely, 4, 096-D convolutional neural networks (CNN) visual features by AlexNet [19], 200-D Stacked Denoising Auto-encoder (SDA) acoustic features, and 100-D paragraph to vector textual features. Noticeably, in our selected dataset, 169 and 24, 707 micro-videos do not have acoustic and textual modalities, respectively. We inferred their missing data via matrix factorization, which have been proven to be effective in the multi-modal data completion task [36].

3.2 External Sound Dataset

As analyzed before, the acoustic modality is the least descriptive one and we expect to borrow the external sounds to enhance its discrimination. The scope of external sound dataset has direct effect on the performance of representation learning over micro-videos. Therefore, external sound construction is of importance. Indeed, there are several prior efforts on the sound clip collection. For example, Mesaros *et al.* [30] manually collected audio recordings from 10 acoustic environments and recognized them into 60 event-oriented concepts; Pancoast *et al.* [33] established 20 acoustic concepts relying on a small subset of TRECVID 2011; Burger *et al.* [3] extracted 42 concepts to describe distinct noise units from the soundtracks of 400 videos. We noticed that the existing external sound bases are either too small to cover the common acoustic concepts, or acquired from a narrow range of event-oriented videos. They are thus infeasible for our task.

To address this problem, we chose to collect sound clips from Freesound. Freesound is a collaborative repository of Creative Commons licensed audio samples with more than 230,000 sounds and 4 million registered users as of February 2015. Short audio clips are uploaded to the website by its users, and cover a wide range of real-life subjects, like *applause* and *breathing*. Audio content in the repository can be tagged with acoustic concepts and browsed by standard text-based search. We first went through a rich set of micro-videos and manually defined 131 acoustic concepts, including the 60 acoustic concepts from the real-life recordings in [29].

Our pre-defined acoustic concepts are diverse and treated as the initial seeds. We then fed these concepts into Freesound as

[6] http://acmmm16.wixsite.com/mm16.

Table 1: Statistics of our collected external sound data.

Concepts #	Total Sound Clip #	Sound Clips # Per Concept	Average Duration	Average Concepts # Per Sound Clip
313	43,868	140.15	14.99 seconds	2.99 (after data laundry)

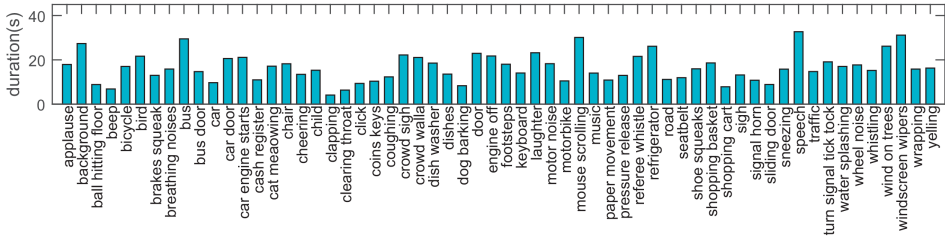


Figure 3: Exemplar demonstration of some acoustic concepts and their average sound clip durations.

queries to search the relevant sound clips. In this way, we gathered 16,363 clips. Each clip was manually labeled with several tags (*i.e.,* acoustic concepts) by their owners and we in total obtained 146,580 acoustic concepts. To select the commonly heard acoustic concepts, we filtered out those concepts with less than 50 sound clips. Meanwhile, we adopted WordNet [18] to merge the acoustic concepts with similar semantic meanings, such as *kids* and *child*. Thereafter, we were left a set of 465 distinct acoustic concepts. Following that, we again fed each acoustic concept into Freesound as a query to acquire its sound clips with a number limit of 500. As a result, we gathered 45,948 sound clips. To ensure the quality of the sound data, we retrained acoustic concepts with at least 100 sound clips. We ultimately have 313 acoustic concepts and 43,868 sound clips. The statistics of the acoustic dataset are summarized in Table 1. Some acoustic concept examples and their average sound durations are demonstrated in Figure 3. We can see that the external sound clips are very short. Similar to the micro-videos, the collected sound clips can be characterized by high-level concepts. Regarding each audio clip, we explored and extracted the same SDA acoustic features with those of the acoustic modality in micro-videos. We will clarify why we extracted this type of features in the experiments.

4 DEEP TRANSFER MODEL

In this section, we formally introduce the problem definition. Suppose there are N micro-videos $\mathcal{X} = \{x_i\}_{i=1}^{N}$. For each micro-video $x \in \mathcal{X}$, we pre-segment it into three modalities $x = \{\mathbf{x}^v, \mathbf{x}^a, \mathbf{x}^t\}$, whereinto the superscripts v, a, and t respectively represents the visual, acoustic, and textual modality. To make more clear presentation, we denote $m \in \mathcal{M} = \{v, a, t\}$ as a modality indicator, and $\mathbf{x}^m \in \mathbb{R}^{D_m}$ as the D_m-dimensional feature vector over the m-th modality. And we associate $\mathbf{x}$ with one of the K pre-defined venue categories, namely an one-hot label vector $\mathbf{y}$. Our research objective is to generalize a venue estimation model over the training set to the new coming micro-videos.

4.1 Sound Knowledge Transfer

In order to leverage external sound knowledge to enhance the acoustic modality in micro-videos, we have two assumptions: 1) Concept-level representations are more discriminative to characterize each modality in micro-videos and the external sounds. And 2) the natural correlation between the acoustic modality in micro-videos and the real-life sounds motivates us to assume that they share the same acoustic concept space.

As to the concept-level representation, one intuitive thought is multi-modal dictionary learning, whereby the atoms in the dictionaries are treated as concepts. We, however, argue that the implicit assumption of multi-modal dictionary learning does not always hold in some real-world scenarios: the dictionaries of distinct modalities share the same concept space. Considering the micro-video as an example, the acoustic modality may contain the concept of *chirp of birds* that is hardly expressed by the visual modality. In the textual one, it may signal some atoms related to *sense of smell*, which also impossibly appear in the visual modality. Therefore, it is not necessary to enforce the dictionaries of different modalities to contain the same set of concepts. To avoid such problem, we propose to learn a separate mapping function for each modality that is able to project the low-level features to concept-level representations. Analogous to the dictionaries in dictionary learning paradigms, the mapping functions are the concept-feature distributions.

Let $\widetilde{\mathcal{X}}^a = \{\widetilde{\mathbf{x}}_i^a\}_{i=1}^{N'}$ be the dataset of external sounds. These sounds share the same low-level feature space with the acoustic modality in micro-videos (*i.e.,* $\widetilde{\mathbf{x}}^a \in \mathbb{R}^{D_a}$). For each sound clip $\widetilde{\mathbf{x}}^a$, we denote its corresponding concept-wise representation as $\widetilde{\mathbf{a}}^a \in \mathbb{R}^{K'}$ over K' acoustic concepts, whereby K' equals to the number of acoustic concepts in this work, *i.e.,* 313. It is worth noting that $\widetilde{\mathbf{a}}^a$ is observable, since we know the associated tags (*i.e.,* acoustic concepts of each collected sound clip). During learning, we aim to use the concept space of the external real-life sounds to represent the acoustic modality in each micro-video. This is accomplished by ensuring that $\mathbf{x}^a$ and $\widetilde{\mathbf{x}}^a$ share the same mapping function. Based upon this, our objective function $\mathcal{J}_1$ of sound knowledge transfer can be stated as:

$$\mathcal{J}_1 = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{m \in \mathcal{M}} \left\| \mathbf{D}^m \mathbf{x}^m - \mathbf{a}^m \right\|^2 + \frac{1}{N'} \sum_{\widetilde{\mathbf{x}} \in \widetilde{\mathcal{X}}} \left\| \mathbf{D}^a \widetilde{\mathbf{x}}^a - \widetilde{\mathbf{a}}^a \right\|^2, \quad (1)$$

where $\mathbf{D}^a \in \mathbb{R}^{D_a \times K'}$ is the shared mapping function, bridging the gap between the external sounds and the internal acoustic modality, whereinto its i-th column $\mathbf{d}_i^a$ represents the low-level feature for the i-th concept, such as *footsteps* or *clearing throat*; and $\mathbf{a}^a \in \mathbb{R}^{K'}$ is the desired concept-level representation of $\mathbf{x}$ over the K' acoustic concepts; $\mathbf{D}^v$ and $\mathbf{a}^v$ ($\mathbf{D}^t$ and $\mathbf{a}^t$) are analogous to $\mathbf{D}^a$ and $\mathbf{a}^a$. Noticeably, $\mathbf{D}^v$ is an identity matrix, slightly different

from other two mapping functions, since the visual features are sufficiently abstractive extracted by AlexNet.

4.2 Multi-modal Fusion

As aforementioned, multi-modalities provide complementary. We thus argue that multi-modal fusion [12, 40] can provide comprehensive and informative description for micro-videos. In our case, we adopt early fusion strategy for simplicity. Formally, for each micro-video $\mathbf{x}$, we concatenate $\mathbf{a}^v$, $\mathbf{a}^a$, and $\mathbf{a}^t$ into one vector as,

$$\mathbf{a} = [\mathbf{a}^v, \mathbf{a}^a, \mathbf{a}^t], \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^{K_v+K'+K_t}$ is the desired multi-modal representation for $\mathbf{x}$, whereinto $\mathbf{a}^v$, $\mathbf{a}^a$, and $\mathbf{a}^t$ respectively denotes the concept-level representation over the visual, acoustic, and textual modalities.

To alleviate the sparsity problem of unpopular categories, we further boost the representation learning of each category by preserving and regularizing the venue similarity. In particular, if two micro-videos are captured in the same venue, they should have similar representations in the latent space; otherwise, their representations should be dissimilar. As such, the representation learning process of unpopular categories can benefit from the processes of other categories since the representation pertains to the discriminative and semantical category information. This suits well the paradigm of graph embedding [40, 43, 45, 46], which injects the label information into the embeddings.

Formally, we denote $(\mathbf{x}_i, \mathbf{x}_j)$ as the pair of the i-th and j-th samples, and define a pairwise class indicator as,

$$\gamma_{ij} = \begin{cases} +1, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ have the same label;} \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

To encode the similarity preservation [17, 39], we minimize the cross entropy loss of classifying all the pairs into a label γ,

$$\sum_{i,j=1}^{N} -\mathbb{I}(\gamma_{ij} = 1) \log \sigma(\mathbf{a}_i^\top \mathbf{a}_j) - \mathbb{I}(\gamma_{ij} = -1) \log \sigma(-\mathbf{a}_i^\top \mathbf{a}_j), \tag{4}$$

where $\mathbb{I}(\cdot)$ is a binary indicator function that outputs 1 when the argument is true, otherwise 0; and $\sigma(\cdot)$ is the sigmoid function. We can equivalently rewrite the above equation as,

$$\mathcal{J}_2 = -\sum_{i=1}^{N} \sum_{j=1}^{N} \log \sigma(\gamma_{ij} \mathbf{a}_i^\top \mathbf{a}_j). \tag{5}$$

It is very time-consuming to directly optimize Eqn.(5) due to the huge amount of the instance pairs, *i.e.,* $O(N^2)$ *w.r.t.* N samples.

To reduce the computing load, we turn to the strategy of negative sampling [31]. In particular, for a given micro-video sample $\mathbf{x}$, we respectively sampled S positive from $\mathbf{x}$'s own category and S negative micro-videos from its non-categories following a distribution $(\mathbf{x}_i, \mathbf{x}_j, \gamma_{ij})$.

4.3 Deep Network for Venue Estimation

After obtaining the multi-modal representations, we add a stack of fully connected layers following [16, 39], which enables us to capture the nonlinear and complex interactions between the visual, acoustic, and textual concepts. More formally, we define these fully connected layers as,

$$\begin{cases} \mathbf{e}_1 = \sigma_1(\mathbf{W}_1 \mathbf{a} + \mathbf{b}_1) \\ \mathbf{e}_2 = \sigma_2(\mathbf{W}_2 \mathbf{e}_1 + \mathbf{b}_2) \\ \cdots\cdots \\ \mathbf{e}_L = \sigma_L(\mathbf{W}_L \mathbf{e}_{L-1} + \mathbf{b}_L) \end{cases}, \tag{6}$$

where $\mathbf{W}_l$, $\mathbf{b}_l$, σ_l, and $\mathbf{e}_l$ denote the weight matrix, bias vector, activation function, and output vector in the l-th hidden layers, respectively. As for activation function in each hidden layer, we choose Rectifier (ReLU) to learn higher-order concept interactions in a non-linear way. Regarding the size of hidden layers, common solutions follow the tower, constant, and diamond patterns.

The output of the penultimate hidden layer is flattened to a dense vector $\mathbf{e}_L$, which is passed to a fully connected softmax layer. It computes the probability distributions over the venue category labels, as,

$$p(\widehat{y}_k | \mathbf{e}_L) = \frac{exp(\mathbf{e}_L^\top \mathbf{w}_k)}{\sum_{k'=1}^{K} exp(\mathbf{e}_L^\top \mathbf{w}_{k'})}, \tag{7}$$

where $\mathbf{w}_k$ is a weight vector of the k-th venue category; $\mathbf{e}_L$ can be viewed as the final abstract representation of the input $\mathbf{x}$. Thereafter, we obtain the probabilistic label vector $\widehat{\mathbf{y}} = [\widehat{y}_1, \cdots, \widehat{y}_K]$ over the K venue categories.

Thereafter, we adopt the regression-based function to minimize the loss between the estimated label vector and its target values, as,

$$\mathcal{J}_3 = \frac{1}{2} \sum_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - \widehat{\mathbf{y}}\|^2, \tag{8}$$

where an ideal model should predict the venue category correctly for each micro-video.

We ultimately obtain our objective function of the proposed deep transfer model by jointly regularizing the sound knowledge transfer, multi-modal fusion, and deep neural network for venue estimation as,

$$\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3. \tag{9}$$

4.4 Training

We adopted the stochastic gradient descent (SGD) to train our model in a mini-batch mode and updated the corresponding model parameters using back propagation. In particular, we first sampled a batch of instances and took a gradient step to optimize the loss function of external sound transfer. We then sampled a batch of $(\mathbf{x}_i, \mathbf{x}_j, \gamma_{ij})$ and took another gradient step to optimize the loss of multi-modal embedding learning. Thereafter, we optimized the loss function of venue category estimation. To speed up the convergence rate of SGD, various modifications to the update rule have been explored, namely, momentum, adagrad, and adadelta.

While deep neural networks are powerful in representation learning, a deep architecture easily leads to the overfitting on the limited training data. To remedy the overfitting issue, we conducted dropout to improve the regularization of our deep model. The idea is to randomly drop part of neurons during training. As such, dropout acts as an approximate model averaging. In particular, we randomly dropped ρ of $\mathbf{a}$, whereinto ρ is the dropout ratio. Analogously, we also conducted dropout on each hidden layer.

5 EXPERIMENTS

To thoroughly justify the effectiveness of our proposed deep transfer model, we carried out extensive experiments to answer the following research questions:

- **RQ1:** Are the extracted 200-D SDA features discriminative to represent the external sounds?
- **RQ2:** Can our DARE approach outperform the state-of-the-art baselines for micro-video categorization?
- **RQ3:** Is the external sound knowledge helpful for boosting the categorization accuracy and does the external data size affect the final results?
- **RQ4:** Does the proposed DARE model converge and do different parameter settings affect the final results?

5.1 Experimental Settings

5.1.1 Metrics. In this work, we adopted Macro-F1 and Micro-F1 [14] to measure the micro-video classification performance of our approach and the baselines. They both reach the best value at 1 and worst one at 0. The macro-average weights all the classes equally, regardless of how many instances belong to each class. By contrast, the micro-average weights all the instances equally.

We divided our dataset into three parts: 132, 370 for training, 56, 731 for validation, and 81, 044 for testing. The training set was used to adjust the parameters, while the validation one was used to verify that any performance increase over the training dataset actually yields an accuracy increase over a dataset that has not been shown to the model before. The testing set was used only for testing the final solution to confirm the actual predictive power of our model with optimal parameters.

5.1.2 Baselines. We chose the following methods as baselines:

- **Default:** For any given micro-video, we dropped it into the category with the most micro-videos by default.
- **D^{3}L:** Data-driven dictionary learning is a classic unimodal supervised dictionary learning framework [26].
- **MDL:** This baseline is the traditional unsupervised multi-modal dictionary learning [32]. It is followed with a softmax classifier.
- **MTDL:** This is a multi-modal task-driven dictionary learning approach [2] learning the discriminative multi-modal dictionaries simultaneously with the corresponding venue category classifiers.
- **TRUMANN:** This is a tree-guided multi-task multi-modal learning method, which considers the hierarchical relatedness among the venue categories.
- **AlexNet:** In addition to the shallow learning methods, we added four deep models into our baseline pool, *i.e.,* the AlexNet model with zero, one, two, and three hidden layers, whereby their inputs are the original feature concatenation of three modalities and they predict the final results with a softmax function.

Indeed, our model is also related to transfer learning methods. However, existing transfer models [10, 23, 44] are not suitable to our task, since they work by leveraging one source domain to support one target domain. Yet, our task has one source domain (external sounds) and three target domains (three modalities). Therefore, we did not compare our method with transfer learning methods.

Table 2: Discrimination comparison among different acoustic features.

Feature sets	Macro-$F1$	Micro-$F1$	p-value
spectrum(mean)	6.23±0.30%	8.87±0.17%	7.4e-6
spectrum(max)	5.88±0.49%	8.25±0.53%	4.8e-6
MFCC(mean)	4.92±0.63%	11.36±0.96%	3.6e-4
MFCC(max)	9.21±0.46%	15.72±0.77%	4.2e-3
SDA	**12.74 ± 0.62%**	**17.09 ± 0.69%**	-

5.1.3 Parameter Settings. We implemented our DARE model with the help of Tensorflow[7]. To be more specific, we randomly initialized the model parameters with a Gaussian distribution for all the deep models in this paper, whereby we set the mean and standard derivation as 0 and 1, respectively. The mini-batch size and learning rate for all models was searched in [256, 512, 1,024] and [0.0001, 0.0005, 0.001, 0.005, 0.1], respectively. We selected Adagrad as the optimizer. Moreover, we selected the constant structure of hidden layers, empirically set the size of each hidden layer as 1,024 and the activation function as ReLU. For our DARE, we set the embedding sizes of visual, acoustic, and textual mapping matrices as 4,096, 313, and 200, respectively, which can be treated as the extra hidden layer for each modality. Without special mention, we employed one hidden layer and one prediction layer for all the deep methods. We randomly generated five different initializations and fed them into our DARE. For other competitors, the initialization procedure is analogous to ensure the fair comparison. We reported the average testing results over five round results and performed paired t-test between our model and each of baselines over five-round results.

5.2 Acoustic Representation (RQ1)

To represent each external sound clip, we fist extracted two kinds of commonly used features, *i.e.,* 513-D spectrum and 39-D mel frequency cepstral coefficients (MFCCs), with a 46-ms window size and 50% overlap via librosa[8]. We then employed the mean- and max-pooling strategy to represent each clip. Besides, we also adopted theano[9] to learn a 200-D SDA feature vector of each clip, whose input is the concatenated feature vector of 513-D spectrum (mean), 513-D spectrum (max), 39-D MFCCs (mean), and 39-D MFCCs (max).

In order to justify the discrimination of the extracted features on the external sounds, we respectively fed the features into a softmax model to learn a sound clip classifier. In particular, we treated each acoustic concept as a label. We performed a 10-fold cross-validation. The results are summarized in Table 2. We can see that the SDA features are the most discriminant one. We conducted significance test between SDA and each of the others regarding Macro-F1 based on the 10-round results. All the p-values are greatly smaller than 0.05, which indicates that SDA is statistically significant better. That is why we used the SDA feature in the hereafter experiments.

5.3 Performance Comparison (RQ2)

We summarized the performance comparison among all the methods in Table 3. We have the following observations:

[7] https://www.tensorflow.org
[8] http://librosa.github.io/librosa.
[9] http://deeplearning.net/software/theano.

Table 3: Performance comparison between our model and the baselines. Thereinto, AlexNet˙L denotes an AlexNet model with L layers.

	Micro-F1	Macro-F1	p-value1*	p-value2*
Default	11.40%	0.53%	1.93e-9	1.41e-8
MDL	20.46±0.49%	7.06±0.27%	3.39e-8	2.01e-7
D^3L	19.03±0.29%	3.87±0.24%	1.29e-8	2.29e-8
MTDL	20.67±0.29%	6.16±0.24%	4.29e-8	1.94e-8
$AlexNet_0$	25.95±0.08%	6.04±0.07%	9.81e-7	1.36e-8
$AlexNet_1$	28.95±0.17%	9.45±0.13%	2.15e-5	1.38e-7
$AlexNet_2$	29.04±0.17%	10.86±0.18%	4.02e-5	1.24e-6
$AlexNet_3$	28.55±0.49%	10.65±0.34%	1.91e-4	4.87e-6
TRUMANN	25.27±0.17%	5.21±0.29%	2.46e-7	9.23e-8
DARE	**31.21 ± 0.22%**	**16.66 ± 0.30%**	-	-

Table 4: Performance of DARE with different hidden layers.

Hidden Layers	Micro-F1	Macro-F1	p-value1	p-value2
[1024]	**31.21±0.22%**	**16.66±0.30%**	-	-
[1024, 1024]	30.67±0.06%	15.57±0.03%	$1.32e-2$	$3.50e-3$
[1024, 1024, 1024]	29.43±0.02%	13.37±0.04%	$1.17e-4$	$1.57e-6$

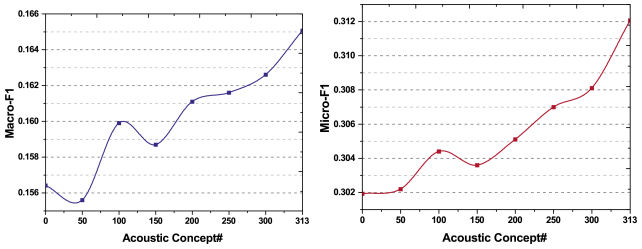
Table 5: Micro-F1 Performance of DARE and DARE-sparsity over the uniformly split venue category groups.

Category IDs	[0-46]	[47-93]	[94-140]	[141-187]
DARE	**34.84 ± 0.32%**	**20.64 ± 0.40%**	**10.34 ± 0.42%**	**7.21 ± 0.34%**
DARE-sparsity	31.75 ± 0.02%	9.63 ± 0.03%	1.53 ± 0.08%	1.67 ± 0.04%

- As expect, Default achieves the worst performance, especially *w.r.t.* Macro-F1.
- In terms of Micro-F1, performance of three dictionary learning baselines is comparative; whereas D^3L achieves the worst Macro-F1. This may be due to that D^3L does not differ the three modalities.
- The TRUMANN model is better than dictionary learning methods, since it considers the hierarchical structure of venue categories.
- AlexNet with at least one hidden layer remarkably outperforms $AlexNet_0$ and dictionary learning ones across metrics. This demonstrates the advantage of deep models.
- Among the AlexNet series, it is not the deeper the better. This is caused by the intrinsic limitation of AlexNet (We will detail it in RQ4.).
- Without a doubt, our proposed model achieves the best regardless of the metrics. This justifies the effectiveness of our model. From the perspective of Macro-F1, our model makes noteworthy progress. This further shows the rationality of similarity preservation by encoding the structural category information. In addition, we also conducted pair-wise significant test between our model and each baseline. All the p-values are greatly smaller than 0.05, which indicates the performance improvement is statistically significant.

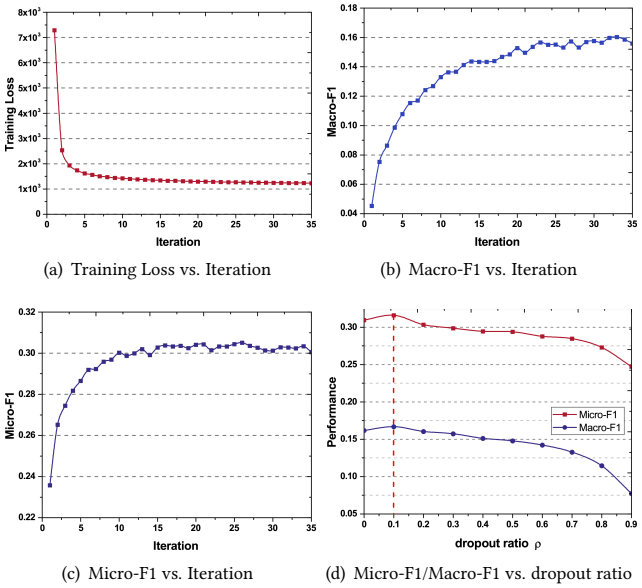
5.4 External Knowledge Effect (RQ3)

We carried out experiments to study the effect of external sound knowledge on our model. In particular, we varied the number of external acoustic concepts from 0 to 313. Figure 4(a) and 4(b) illustrates the performance of our model according to the external data size *w.r.t.* Macro-F1 and Micro-F1, respectively. It is clear that these two curves goes up very fast. Such phenomenons tell



(a) Macro-F1 vs. Acoustic Concept (b) Micro-F1 vs. Acoustic Concept

Figure 4: Performance of DARE *w.r.t.* the number of external acoustic concepts in terms of F1 measurements.



(a) Training Loss vs. Iteration (b) Macro-F1 vs. Iteration

(c) Micro-F1 vs. Iteration (d) Micro-F1/Macro-F1 vs. dropout ratio

Figure 5: Convergence and dropout ratio study of the proposed DARE model.

us that transferring external sound knowledge is useful to boost the categorization accuracy. Also, it signals that the more external sounds are involved, the better performance we will achieve. This is because it can cover a much wider range of acoustic concepts appeared in micro-videos, and hence better strengthen the acoustic modality comprehensively.

5.5 Study of DARE Model (RQ4)

We wonder whether our model converges and how fast it is. To answer this question, we plot the training loss, Macro-F1, and Micro-F1 *w.r.t.* the number of iterations in Figure 5(a), 5(b), and 5(c), respectively. From these three sub-figures, it can be seen that the training loss of our proposed DARE model decreases quickly within the first 10 iterations, and accordingly the performance is also boosted very fast. This demonstrates the rationality of a learning model. In addition, the loss and performance tend to be stable at around 30 iterations. This signals the convergence property of our model and also indicates its efficiency.

The key idea of dropout technique is to randomly drop units (along with their connections) from the neural network during

training. This prevents units from co-adapting too much. Figure 5(d) displays the Macro-F1 and Micro-F1 by varying the dropout ratio ρ. From this figure, it can be seen that the two measurements consistently reach their best value when using a dropout ratio of 0.1. After 0.1, the performance decreases gradually as the dropout ratio increases. This may be caused by insufficient information. Also, we can see that our model suffers from overfitting with relatively lower performance when dropout ratio is set as 0.

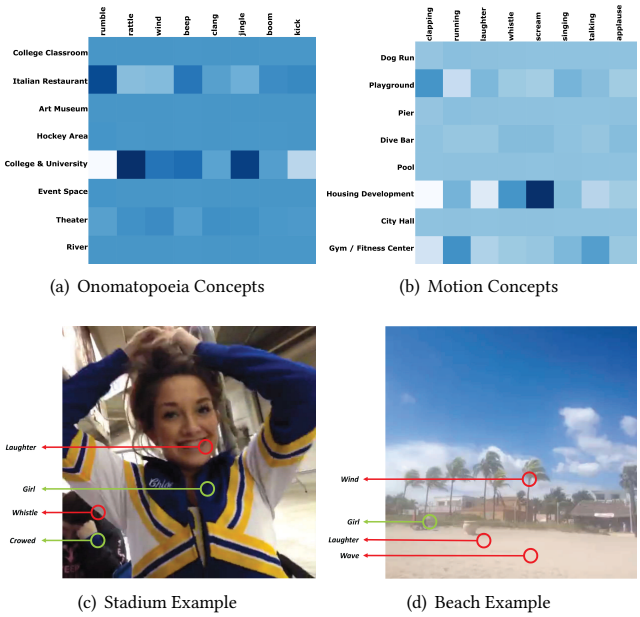
We also studied the impact of hidden layers on our DARE model. To save the computational tuning costs, we applied the same dropout ratio 0.1 for each hidden layer. The results of our model with different hidden layers are summarized in Table 4. Usually, stacking more hidden layers is beneficial to boost the desired performance. However, we notice that our model achieves the best across metrics when having only one hidden layer. This is due to that, as the authors of AlexNet clarified, the current 7-layer AlexNet structure is optimal and more layers would lead to worse results. Therefore, stacking more hidden layers in our DARE model seems to add more hidden layers to AlexNet.

To analyze the effect of the proposed similarity regularizer, we remove the Eqn.(5) from our DARE model, denoted as DARE-sparsity. Thereafter, we conduct DARE and DARE-sparsity over the same venue category groups, where the smaller category IDs represent the more popular venues with more sufficient training data. Table 5 presents the performance *w.r.t.* Micro-F1 over different category groups. As we can see, when category groups tend to be unpopular, DARE-sparsity suffers severely from the insufficient training data and the poor classifier; meanwhile, DARE is relatively insensitive to the sparsity problem and can boost the classification performance considerably. This highlights the significance of similarity regularization of categories.

5.6 Visualization

We conducted experiments to shed some light on the correlation between venue categories and acoustic concepts. In particular, we calculated the correlations between acoustic concepts and venue categories via producing inner products on the conceptual distributions and venue label vectors of samples.

- To save the space, we visualized part of correlation matrix via a heat map, where lighter color indicates weak correlation and vice versa, as shown in Figures 6(a) and 6(b). We can see that almost every selected venue category are tightly related to several acoustic concepts. Moreover, different venues emphasize a variety of acoustic concepts. For example, the micro-videos with venue of *Italian Restaurant* and *College (University)* have significant correlations with the onomatopoeia concepts, such as *rattle*, *jingle*, and *rumble*; Meanwhile, several motion concepts, such as *screaming*, *running*, and *clapping* provide clear cues to infer the venue information of *Housing Development*, *Gym*, and *Playground*, respectively. These observations agree with our common sense and further demonstrates the potential influence of acoustic information on the task of venue category estimation.
- We select exemplary demonstration of two micro-videos of stadium and beach, as shown in Figures 6(c) and 6(d). The detected acoustic concepts include several concepts which are consistent with the visual modality, such as *girl* and *crowed*, and some exclusive ones which are hardly revealed from the other modalities, such as *laughter*, *whistle*, and *wave*. It further verifies our assumption that the concepts from different modalities are complementary with each other.



(a) Onomatopoeia Concepts (b) Motion Concepts

(c) Stadium Example (d) Beach Example

Figure 6: (a)-(b): Visualization of correlations between venue category and two types of acoustic concepts; (c)-(d): Exemplary demonstration of two micro-videos of stadium and beach, where the red circles present the detected acoustic concepts that are hardly detected from the visual modality, and the green circles denote these concepts that are consistent with the visual modality. (It is noted that the positions of circle do not present the locations of concepts).

6 CONCLUSION AND FUTURE WORK

In this paper, we study the task of micro-video category estimation. We present a deep transfer model, which is able to transfer external sound knowledge to strengthen the low-quality acoustic modality in micro-videos, and also alleviate the sparsity problem via encoding the category information into the representation learning. To justify our model, we constructed the external sound sets with diverse acoustic concepts, and released it to facilitate the community research. Experimental results on a public benchmark micro-video dataset well validated our model.

In the future, we plan to avoid the AlexNet limitation by exploring ResNet framework. Furthermore, pointing out the shared and exclusive concepts of various modalities is our next research focus, rather than simply detecting the concepts. Moreover, we will introduce the attention scheme into our work to explicitly estimate the influence of each concept on different venues.

Acknowledgement We would like to thank the anonymous reviewers for their valuable comments. The work is supported by the One Thousand Talents Plan of China under Grant No.:11150087963001.

REFERENCES

[1] Khalid Ashraf, Benjamin Elizalde, Forrest Iandola, Matthew Moskewicz, Julia Bernd, Gerald Friedland, and Kurt Keutzer. 2015. Audio-based multimedia event detection with DNNs and sparse sampling. In *ICMR*. 611–614.

[2] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins. 2016. Multimodal task-driven dictionary learning for image classification. *TIP* 25, 1 (2016), 24–38.

[3] Susanne Burger, Qin Jin, Peter F Schulam, and Florian Metze. 2012. Noisemes: Manual annotation of environmental noise in audio streams. *Technical report CMU-LTI-12-07* (2012), 1–5.

[4] Song Cao and Noah Snavely. 2013. Graph-based discriminative learning for location recognition. In *CVPR*. 700–707.

[5] Diego Castan and Murat Akbacak. 2013. Segmental-GMM Approach based on Acoustic Concept Segmentation. In *SLAM@ INTERSPEECH*. 15–19.

[6] Sourish Chaudhuri and Bhiksha Raj. 2012. Unsupervised structure discovery for semantic analysis of audio. In *NIPS*. 1178–1186.

[7] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro Tells Macro: Predicting the Popularity of Micro-Videos via a Transductive Model. In *MM*. 898–907.

[8] Ning Chen, Jun Zhu, and Eric P Xing. 2010. Predictive subspace learning for multi-view data: a large margin approach. In *NIPS*. 361–369.

[9] Jaeyoung Choi, Gerald Friedland, Venkatesan Ekambaram, and Kannan Ramchandran. 2012. Multimodal location estimation of consumer media: Dealing with sparse training data. In *ICME*. 43–48.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*. 647–655.

[11] M. Elad and M. Aharon. 2006. Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *TIP* 15, 12 (2006), 3736–3745.

[12] Fuli Feng, Liqiang Nie, Xiang Wang, Richang Hong, and Tat-Seng Chua. 2017. Computational social indicators: a case study of Chinese university ranking. In *SIGIR*.

[13] Gerald Friedland, Jaeyoung Choi, Howard Lei, and Adam Janin. 2011. Multimodal location estimation on Flickr videos. In *MM*. 23–28.

[14] Siddharth Gopal and Yiming Yang. 2013. Recursive Regularization for Large-scale Classification with Hierarchical and Graphical Dependencies. In *SIGKDD*. 257–265.

[15] James Hays and Alexei A Efros. 2008. IM2GPS: estimating geographic information from a single image. In *CVPR*. 1–8.

[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*.

[17] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Chua Tat-Seng. 2016. Fast Matrix Factorization for Online Recommendation with Implicit Feedback. In *SIGIR*.

[18] Adam Kilgarriff and Christiane Fellbaum. 2000. WordNet: An Electronic Lexical Database. (2000).

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. 1106–1114.

[20] Anan Liu, Weizhi Nie, Yue Gao, and Yuting Su. 2016. Multi-Modal Clique-Graph Matching for View-Based 3D Model Retrieval. *TIP* 25, 5 (2016), 2103–2116.

[21] Anan Liu, Yuting Su, Weizhi Nie, and Mohan S. Kankanhalli. 2017. Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition. *TPAMI* 39, 1 (2017), 102–114.

[22] Gaowen Liu, Yan Yan, Elisa Ricci, Yi Yang, Yahong Han, Stefan Winkler, and Nicu Sebe. 2015. Inferring Painting Style with Multi-task Dictionary Learning. In *IJCAI*. 2162–2168.

[23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *ICML*. 97–105.

[24] J. Mairal, F. Bach, and J. Ponce. 2012. Task-Driven Dictionary Learning. *TPAMI* 34, 4 (2012), 791–804.

[25] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online Dictionary Learning for Sparse Coding. In *ICML*. 689–696.

[26] Julien Mairal, Francis R. Bach, and Jean Ponce. 2012. Task-Driven Dictionary Learning. *TPAMI* 34, 4 (2012), 791–804.

[27] Julien Mairal, Michael Elad, and Guillermo Sapiro. 2008. Sparse representation for color image restoration. *TIP* 17, 1 (2008), 53–69.

[28] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R. Bach. 2009. Supervised Dictionary Learning. In *NIPS*. 1033–1040.

[29] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *EUSIPCO*. 1267–1271.

[30] Annamaria Mesaros, Toni Heittola, Antti J. Eronen, and Tuomas Virtanen. 2010. Acoustic event detection in real life recordings. In *EUSIPCO*. 1267–1271.

[31] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*. 3111–3119.

[32] Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhe, Sylvain Lesage, and Rémi Gribonval. 2007. Learning multimodal dictionaries. *TIP* 16, 9 (2007), 2272–2283.

[33] Stephanie Lynne Pancoast, Murat Akbacak, and Michelle Hewlett Sanchez. 2012. Supervised acoustic concept extraction for multimedia event detection. In *Proceedings of the 2012 ACM international workshop on Audio and multimedia methods for large-scale video analysis*. ACM, 9–14.

[34] Mirco Ravanelli, Benjamin Elizalde, Karl Ni, and Gerald Friedland. 2014. Audio concept classification with hierarchical deep neural networks. In *EUSIPCO*. 606–610.

[35] S. Sadanand and J. J. Corso. 2012. Action bank: A high-level representation of activity in video. In *CVPR*. 1234–1241.

[36] Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR*. 213–222.

[37] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Guo-Jun Qi, and Yan Song. 2009. Unified video annotation via multigraph learning. *TCSVT* 19, 5 (2009), 733–746.

[38] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012. Multimodal graph-based reranking for web image search. *TIP* 21, 11 (2012), 4649–4661.

[39] Xiang Wang, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2017. Item Silk Road: Recommending Items from Information Domains to Social Users. (2017).

[40] Xiang Wang, Liqiang Nie, Xuemeng Song, Dongxiang Zhang, and Tat-Seng Chua. 2017. Unifying virtual and physical worlds: Learning toward local and global consistency. *TOIS* 36, 1 (2017), 4.

[41] Yipei Wang, Shourabh Rawat, and Florian Metze. 2014. Exploring audio semantic concepts for event-based video retrieval. In *ICASSP*. 1360–1364.

[42] Meng Yang, Weiyang Liu, Weixin Luo, and Linlin Shen. 2016. Analysis-Synthesis Dictionary Learning for Universality-Particularity Representation Based Classification. In *AAAI*. 2251–2257.

[43] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2016. Revisiting Semi-Supervised Learning with Graph Embeddings. In *ICML*. 40–48.

[44] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *NIPS*. 3320–3328.

[45] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *CVPR*.

[46] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, and Tat-Seng Chua. 2014. Robust (semi) nonnegative graph embedding. *TIP* 23, 7 (2014), 2996–3012.

[47] Jianglong Zhang, Liqiang Nie, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2016. Shorter-is-Better: Venue Category Estimation from Micro-Video. In *MM*. 1415–1424.

[48] Yueting Zhuang, Yanfei Wang, Fei Wu, Yin Zhang, and Weiming Lu. 2013. Supervised Coupled Dictionary Learning with Group Structures for Multi-modal Retrieval. In *AAAI*. 1070–1076.