Hierarchy-Dependent Cross-Platform Multi-View Feature Learning for Venue Category Prediction

Shuqiang Jiang, *Senior Member, IEEE*, Weiqing Min, *Member, IEEE*, and Shuhuan Mei

***Abstract*—In this work, we focus on visual venue category prediction, which can facilitate various applications for location-based service and personalization. Considering that the complementarity of different media platforms, it is reasonable to leverage venue-relevant media data from different platforms to boost the prediction performance. Intuitively, recognizing one venue category involves multiple semantic cues, especially objects and scenes, and thus they should contribute together to venue category prediction. In addition, these venues can be organized in a natural hierarchical structure, which provides prior knowledge to guide venue category estimation. Taking these aspects into account, we propose a Hierarchy-dependent Cross-platform Multi-view Feature Learning (HCM-FL) framework for venue category prediction from videos by leveraging images from other platforms. HCM-FL includes two major components, namely Cross-Platform Transfer Deep Learning (CPTDL) and Multi-View Feature Learning with the Hierarchical Venue Structure (MVFL-HVS). CPTDL is capable of reinforcing the learned deep network from videos using images from other platforms. Specifically, CPTDL first trained a deep network using videos. These images from other platforms are filtered by the learnt network and these selected images are then fed into this learnt network to enhance it. Two kinds of pre-trained networks on the ImageNet and Places dataset are employed. Therefore, we can harness both object-oriented and scene-oriented deep features through these enhanced deep networks. MVFL-HVS is then developed to enable multi-view feature fusion. It is capable of embedding the hierarchical structure ontology to support more discriminative joint feature learning. We conduct the experiment on videos from Vine and images from Foursqure. These experimental results demonstrate the advantage of our proposed framework in jointly utilizing multi-platform data, multi-view deep features and hierarchical venue structure knowledge.**

I. Introduction

Recently, visual geo-localization has received a significant amount of attention in both computer vision and multimedia community [34], [37], [25], [42], [32], [3] because of its various applications, such as location based recommendation service [41], augmented reality [16] and photo forensics[1]. One task of visual geo-localization is visual venue category prediction and its goal is to predict the venue category (e.g., Dessert Shop and Pet Store) from images or videos. It is especially important in social media applications such as venue recommendation [41] and tourist route planning. For example we could recommend him/her a particular venue (e.g., Chinese restaurant and movie theater) and give more accurate check-in suggestions based on predicted venue categories a user has visited. Therefore, in this work, we focus on visual venue category prediction.

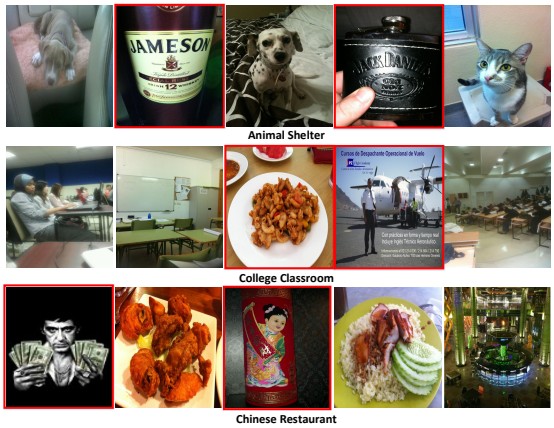

Fig. 1: Some venue categories in Foursquare, where the images labeled with red boxes are noisy ones.

There are some previous works on venue category prediction. For example, Chen *et al.* [1] mined business-aware visual concepts from social media to recognize the business venue from images. Zhang *et al.* [42] proposed a tree-guided multi-task multi-modal learning approach to jointly fuse multi-modal information from videos for venue category prediction. Recently, Nie *et al.* [25] used external sound knowledge to enhance the acoustic modality for venue category estimation from videos. These works mainly utilized information from a single platform for this task. However, little work has investigated this problem via exploiting media data from other platforms, which is especially vital in the deep learning era.

With the success of photo-sharing social websites, we can easily crawl sufficient venue-annotated images from various platforms, such as Foursquare and Instagram. Meanwhile, motivated by the promising results of deep networks on visual analysis tasks, there have also been a number of attempts to utilize deep networks for venue category recognition from videos [25], [42]. However, training such deep networks generally need large-scale data. Therefore, automatically sampling more image data from other platforms appears as a natural

S. Jiang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and also with University of Chinese Academy of Sciences, Beijing, 100049, China email: sqjiang@ict.ac.cn. W. Min is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and also with State key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China. email:minweiqing@ict.ac.cn. S. Mei is with Shandong University of Science and Technology, Shandong, 266590, China, and also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China shuhuan.mei@vipl.ict.ac.cn.

[1]https://traffickcam.com/about

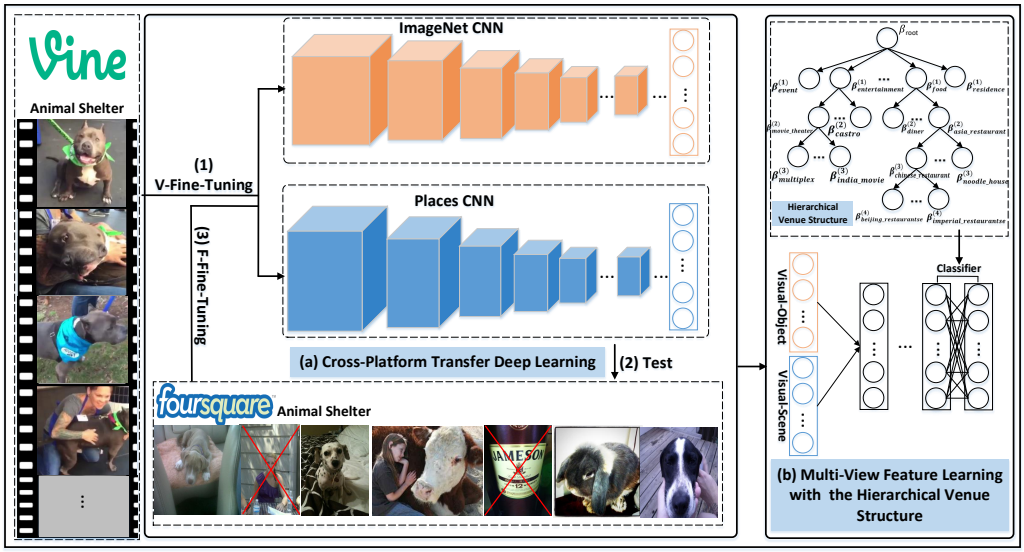

Fig. 2: The proposed Hierarchy-dependent Cross-platform Multi-view Feature Learning (HCM-FL) framework.

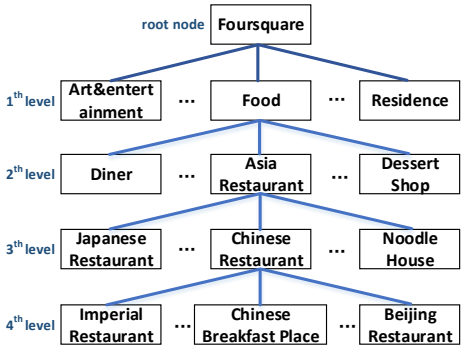

Fig. 3: The hierarchical structure of the venue categories in Foursquare. We illustrate a part of the structure due to the limited space.

way to improve the prediction performance. For example, some works [8], [7] have leveraged web images and videos from Google and Youtube for video classification. However, directly using images from other platforms probably hurts the performance for the following two reasons. First, these images from social websites are noisy. Fig. 1 shows some examples of venue categories from Foursquare. We can see that there are many noisy ones, and the content of some images is irrelevant to the labeled venues. In addition, there is a well-known domain gap problem [27] for the data from different platforms.

Intuitively, recognizing a venue category involves various semantic cues, especially objects and scenes. For example, if there is a dog in one video, then some venue categories such as "Animal Shelter" and "Pet Store" become a probable one. If one video is about the "indoor scene", then the probability of some venue categories such as "Monument" and "Road" reduces. Therefore, both scene and object semantics provide strong context for venue category prediction. Another point to consider is that there are sometimes natural hierarchical structures for venues. For example, the venues in Vine[2] are organized in a four-layer tree structure[3]. Fig. 3 shows a small part of this structure. Knowing the venue structure allows us to borrow the knowledge from relevant venue categories to learn more discriminative features, especially for venue categories with less samples. For example, it is likely that these uploaded videos in the Japanese Restaurant are more similar to ones in the Seafood Restaurant than those in the Christmas Market. In addition, the prediction performance is affected by unbalanced samples on different venue categories. For example, there might be many videos for the theme park but less ones for the Christmas Market in Vine.

Taking all the above-mentioned factors into consideration, we propose a Hierarchy-dependent Cross-platform Multi-view Feature Learning (HCM-FL) framework for venue category prediction from videos. As shown in Fig. 2, HCM-FL mainly consists of two components: (a) Cross-Platform Transfer Deep Learning (CPTDL) (b) Multi-View Feature Learning with the Hierarchical Venue Structure (MVFL-HVS). In particular, we take two platforms Vine and Foursquare in our study and focus on predicting venue categories from videos in Vine by exploiting images from Foursquare. CPTDL first trained a deep network using videos in Vine. The images from Foursquare are filtered by the learnt network and these selected images are then fed into this learnt network to enhance it. Two kinds of pre-trained networks are employed based on ImageNet1000 [20] and Places205 [46]. Therefore, we can harness both object-oriented and scene-oriented deep features through these two kinds of deep networks enhanced by CPTDL, respectively. MVFL-HVS is then developed to learn joint representation from these two types of features. Furthermore, MVFL-HVS can embed the hierarchical structural ontology among venue categories to make learned joint features more discriminative.

The contributions of our paper can be summarized as

[2]The associated venues of the videos in vine are mapped to venue categories in Foursquare.

[3]https://developer.foursquare.com/docs/resources/categories

follows:

- To our knowledge, this is the first study of cross-platform based venue category prediction from videos, where we utilized the transfer deep learning method to enhance the trained network from videos in one platform by taking full advantage of images from the other platform.
- We proposed a hierarchy-dependent cross-platform multi-view feature learning framework for venue category prediction. In this framework, we further developed a multi-view feature learning network, which can embed the hierarchical venue structure knowledge to enable more discriminative joint feature learning.
- We conducted the experiment on two platforms Vine and Foursquare, and these experimental results validated the effectiveness of our proposed framework in fully utilizing multi-platform data, object-scene semantic features and hierarchical venue structure knowledge.

II. The Proposed Framework

As shown in Fig. 2, in this section, we introduce our proposed Hierarchy-dependent Cross-platform Multi-view Feature Learning (HCM-FL) framework, which mainly consists of two components, namely (a) Cross-Platform Transfer Deep Learning (CPTDL) and (b) Multi-View Feature Learning with the Hierarchical Venue Structure (MVFL-HVS). As mentioned before, our task is to predict venue categories from videos in Vine by exploiting images from Foursquare. CPTDL first uses images and videos from two platforms to reinforce the training on two kinds of pre-trained deep networks, namely ImageNet CNN and Places CNN, respectively. Based on two kinds of enhanced networks, we harness both object-oriented and scene-oriented deep visual features for each video. MVFL-HVS is then developed to fuse these two kinds of deep features into a unified feature representation. By embedding the hierarchical venue structure, the fused features are more discriminative. We next introduce each component in details.

A. Cross-Platform Transfer Deep Learning (CPTDL)

As shown in Fig. 2(a), we adopt the VGG-16 deep network [29] as the basic architecture in CPTDL. Particularly, we use two kinds of pre-trained VGG-16 networks: ImageNet CNN, pre-trained on the ImageNet1000 dataset [20], and Places CNN, pre-trained on the Places205 dataset [46]. ImageNet CNN is mainly used to extract the visual object features while Places CNN is mainly used to extract visual scene features [17]. They are complementary and contribute together to venue category prediction.

CPTDL can enhance ImageNet CNN and Places CNN, respectively. We take the ImageNet CNN as an example to describe the training process of CPTDL. Since our task is to predict venue categories from videos in Vine, we start by training a network using the training set of videos. For this training, each video is decomposed into a set of key frames. We first use these key frames to fine-tune the ImageNet CNN, that is V-Fine-Tuning. Then this fine-tuned network is used to test images from Foursquare to filter out noisy images. By utilizing the remaining images from Foursquare to further fine-tune this network, namely F-Fine-Tuning, we can obtain the final enhanced ImageNet CNN.

Note that after V-Fine-Tuning, we next should utilize the images from Foursquare to improve the fine-tuned ImageNet CNN. Although related Foursquare images are helpful for venue prediction, there are usually noisy ones in Foursquare. Take the class "Animal Shelter" as an example (Fig. 2(a)), some images do not describe this venue category. In order to remove useless Foursquare images and keep related ones, we use the ImageNet CNN after V-Fine-Tuning to perform filtering. Formally, for (x_m, t_m), where x_m denotes the m-th image from Foursquare and $t_m \in T$ is its category label. $|T|$ is the number of venue categories. Each image x_m is fed into the fine-tuned ImageNet CNN in a feed-forward way, and yields a probability distribution $p_m \in R^{|T|}$ over the $|T|$ video venue categories. We use $p_m(t)$ to denote the probability of image m belonging to the t-th category. We keep the image m as the t_m-th category if $p_m(t_m)$ is in the top K-ranked probability, where K is a threshold, and $K = 100$ in our experiment. The cleaned Foursquare images are then used to further fine-tune the ImageNet CNN and obtain the final network, which focuses more on video venue categories enhanced by cleaned Foursquare images.

We can adopt a similar training process to obtain the enhanced Places CNN. We then extract visual object features and visual scene features for key frames from training videos based on the enhanced ImageNet CNN and Places CNN, respectively.

B. Multi-View Feature Learning with the Hierarchical Venue Structure(MVFL-HVS)

Through CPTDL, for each key frame from each training video, we obtain visual object features and scene features from enhanced ImageNet CNN and Places CNN, respectively. In order to fuse these two kinds of features, we first transform visual features from key frames of videos to the video-level feature representation. Particularly, for visual object features, we use the output of the FC7 layer from the enhanced ImageNet CNN as the input to the fusion network. That is, for the j-th key frame of video i, $\mathbf{f}_{i,j}$, this pathway outputs $\mathbf{f}_{i,j} \mapsto \mathbf{x}^O_{i,j} \in \mathbb{R}^{4096}$. After mean pooling, we obtain the feature representation $\bar{\mathbf{x}}^O_i = \sum_{j=1}^{n_i} x^O_{i,j}$, where n_i is the number of key frames for video i. We adopt a similar strategy and use the output of the FC7 layer from the enhanced Places CNN to obtain visual scene features $\bar{\mathbf{x}}^S_i = \sum_{j=1}^{n_i} x^S_{i,j}$ for each training video. These two kinds of features are further used as the input to learn a multi-view feature fusion network via MVFL-HVS (Fig. 2(b)).

Different from existing supervised multi-view deep feature fusion networks, MVFL-HVS can embed the hierarchical venue structure ontology to support more discriminative joint feature learning. The venues of videos in Vine are organized in a hierarchical way with the four-layer ontology (Fig. 3). These Hierarchical Venue Structures (HVS) can be used to guide the venue category prediction. As shown in Fig.2(b), we next introduce how to utilize the hierarchical structure prior into our multi-view feature learning network via MVFL-HVS.

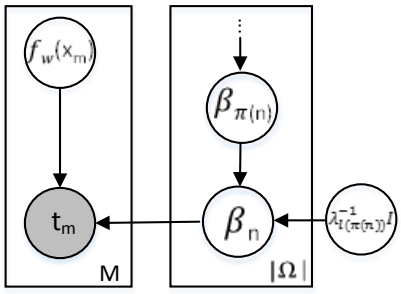

Fig. 4: The graphical representation of the hierarchical structure model.

We define a hierarchy as a set of nodes $\Omega = \{1, 2, ...\}$ with the parent relationship $\pi : \Omega \to \Omega$, where $\pi(n)$ is the parent of node $n \in \Omega$. C_n is the set of all the children of node n. $D = \{(f_{\mathbf{w}}(\mathbf{x}_m), t_m)\}_{m=1}^{M}$ denotes the training data, where x_m is an instance, $f_{\mathbf{w}}(\mathbf{x}_m) \in \mathbb{R}^d$ is the transformed representation through our multi-view fusion network with parameters $\mathbf{w}$. In our work, a multi-layer feedforward neural network is adopted as the fusion network. $t_m \in T$ is its venue category. d is the dimension of transformed features. $T \in \Omega$ is the set of leaf nodes in the hierarchy labeled from 1 to $|T|$. M is the size of training set. We assume that each instance is assigned to one of leaf nodes in the hierarchy. The top-level weight parameters of the last fully-connected classifier layer from the fusion network is$\{\boldsymbol{\beta}_n\}_{n=1}^{|T|}$, where $\boldsymbol{\beta}_n \in \mathbb{R}^{d\times 1}$.

The parent-child relationship is modeled by placing a hierarchical prior over the children nodes centered around the parameters of their parents. Therefore, it can encourage venue categories nearby in the hierarchy to share similar model parameters. The hierarchical graph representation is shown in Fig. 4 and the joint probability distribution is

$$p(\mathbf{D}, \boldsymbol{\beta}, \mathbf{w}) = \prod_m p(t_m|f_{\mathbf{w}}(\mathbf{x}_m), \{\boldsymbol{\beta}_n\}_{n\in T}) \prod_{n\in\Omega} p(\boldsymbol{\beta}_n|\boldsymbol{\beta}_{\pi(n)}, \lambda^{-1}_{l(\pi(n))}\mathbf{I}) \quad (1)$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta}_n\}_{n\in\Omega}$. $\lambda^{-1}_{l(\pi(n))}I$ is a diagonal covariance matrix with the diagonal element $\lambda^{-1}_{l(\pi(n))}$. $\lambda_{l(\pi(n))}$ are hyper-parameters. $l(\pi(n))$ is a mapping function, which maps the index of the node to the corresponding layer/level.

$$p(\boldsymbol{\beta}_n|\boldsymbol{\beta}_{\pi(n)}, \lambda^{-1}_{l(\pi(n))}\mathbf{I}) = \mathcal{N}(\boldsymbol{\beta}_n|\boldsymbol{\beta}_{\pi(n)}, \lambda^{-1}_{l(\pi(n))}\mathbf{I}) \quad (2)$$

where $\mathcal{N}(\cdot)$ is the Gaussian distribution with the mean $\boldsymbol{\beta}_{\pi(n)}$ and the covariance matrix $\lambda^{-1}_{l(\pi(n))}\mathbf{I}$.

The posterior distribution is $p(\boldsymbol{\beta}, \mathbf{w}|\mathbf{D}) \propto p(\mathbf{D}, \boldsymbol{\beta}, \mathbf{w})$. We then use the Maximum A Posteriori probability (MAP) to estimate parameters $\{\mathbf{w}, \boldsymbol{\beta}\}$.

$$\begin{aligned}\log p(\boldsymbol{\beta}, \mathbf{w}|\mathbf{D}) \propto & \sum_m \log p(t_m|f_{\mathbf{w}}(\mathbf{x}_m), \{\boldsymbol{\beta}_n\}_{n\in T}) \\ & + \sum_{n\in\Omega}[-\frac{d}{2}\log 2\pi + \frac{d}{2}\log\lambda_{l(\pi(n))}] \\ & - \sum_{n\in\Omega}[\frac{\lambda_{l(\pi(n))}}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})^\top(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})]\end{aligned} \quad (3)$$

That is, we should minimize

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \mathbf{w}) = & -\sum_m \log p(t_m|f_{\mathbf{w}}(\mathbf{x}_m), \{\boldsymbol{\beta}_n\}_{n\in T}) \\ & + \sum_{n\in\Omega}[\frac{\lambda_{l(\pi(n))}}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})^\top(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})]\end{aligned} \quad (4)$$

where the softmax classifier is used.

The loss function $\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})$ in Eq. 4 can be optimized by iteratively performing the following two steps.

(i) Minimizing over $\{\boldsymbol{\beta}_n\}_{n\in T}$ and $\mathbf{w}$ keeping $\{\boldsymbol{\beta}_n\}_{n\in\Omega\setminus T}$ fixed. This can be implemented using standard Stochastic Gradient Descent (SGD) algorithm to minimize

$$\begin{aligned}\min_{\mathbf{w}, \{\boldsymbol{\beta}_n\}_{n\in T}} \{ & -\sum_m p(t_m|f_{\mathbf{w}}(\mathbf{x}_m), \{\boldsymbol{\beta}_n\}_{n\in T}) \\ & + \sum_{n\in T} \frac{\lambda_{l(\pi(n))}}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})^\top(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})\}\end{aligned} \quad (5)$$

(ii) Minimizing over $\{\boldsymbol{\beta}_n\}_{n\in\Omega\setminus T}$ keeping $\{\boldsymbol{\beta}_n\}_{n\in T}$ and $\mathbf{w}$ fixed. In this step, we should minimize

$$\min_{\{\boldsymbol{\beta}_n\}_{n\in\Omega\setminus T}} \sum_{n\in\Omega\setminus T} \frac{\lambda_{l(\pi(n))}}{2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)})^\top(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)}) \quad (6)$$

When $n \in \Omega\setminus T$,

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})}{\partial\boldsymbol{\beta}_n} = -\sum_{c\in C_n} \lambda_{l(n)}(\boldsymbol{\beta}_c - \boldsymbol{\beta}_n) + \lambda_{l(\pi(n))}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{\pi(n)}) \quad (7)$$

Let $\frac{\partial\mathcal{L}(\boldsymbol{\beta}, \mathbf{w})}{\partial\boldsymbol{\beta}_n} = 0$, we obtain

$$\boldsymbol{\beta}_n = \frac{\sum_{c\in C_n}\lambda_{l(n)}\boldsymbol{\beta}_c + \lambda_{l(\pi(n))}\boldsymbol{\beta}_{\pi(n)}}{\sum_{c\in C_n}\lambda_{l(n)} + \lambda_{l(\pi(n))}} \quad (8)$$

During a training epoch, the forward pass will generate the input $f_{\mathbf{w}}(\mathbf{x}_m)$ for our own loss layer. We then optimize $\{\boldsymbol{\beta}_n\}_{n\in\Omega}$ and $\mathbf{w}$ iteratively according to (i) and (ii). They are then taken back in the backward pass alongside the gradients with respect to its input.

Once the MVFL-HVS is trained, we can predict the venue category of videos in the test stage. After obtaining key frames from one test video, we first extract object features and scene features based on the ImageNet CNN and Places CNN enhanced by CPTDL. We then obtain each kind of video features via mean-pooling on features from key frames. These two kinds of video features are then fed into the trained multi-view feature fusion network from MVFL-HVS to obtain its probability distribution on all the venue categories. The venue with the highest probability is selected as the venue category of this test video.

III. Experiment

A. Dataset

Vine Dataset. We use the dataset with 270,145 micro-videos from [42]. Each video is about 6 seconds in length. Many videos are automatically aligned with a venue category from Foursquare. There are totally 188 venue categories in this dataset. Foursquare organizes its venue categories into a four-layer hierarchical structure with 10, 389, 314 and 52 nodes

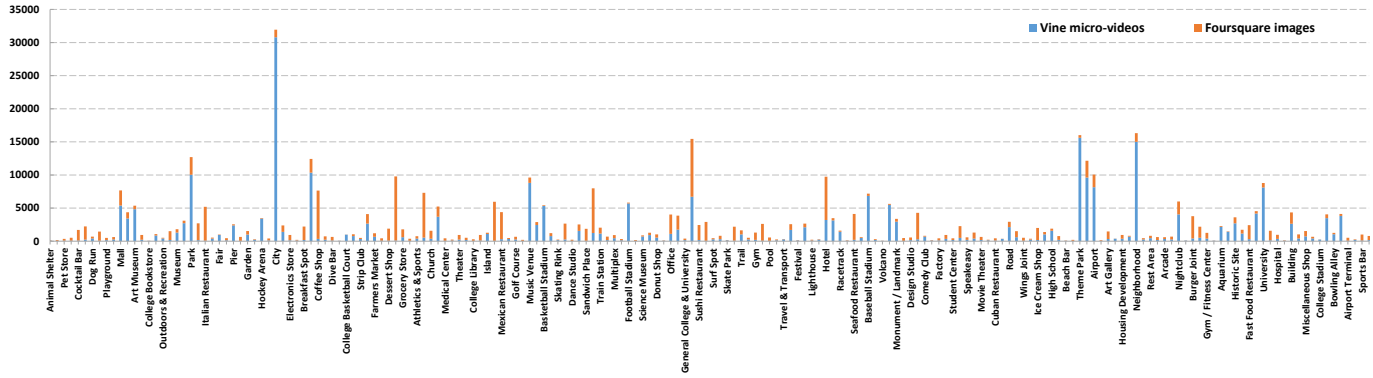

Fig. 5: The statistics of samples for 188 venue categories in two platforms (best viewed under magnification).

Fig. 6: Some example samples from Vine and Foursquare. Each row has 4 videos from Vine and 4 images from Foursquare with the same venue labels.

in the first-layer, second-layer, third-layer and fourth-layer, respectively. Fig. 3 shows a part of the hierarchical structure. Most of venue categories in this dataset are in the third layer, and the remaining ones are in the second layer.

Foursquare Dataset. We use the Foursquare images from [23], where each image is associated with one venue label. We download images using the provided urls and discard records for which venue labels do not belong to these 188 venue categories. The resulting Foursquare dataset consists of 190,299 images.

Fig. 5 provides statistics on both videos from Vine and images from Foursquare while Fig. 6 shows some videos and images from three venue categories.

B. Implementation Details

In order to extract features from videos, similar to [42], we first select key frames from each video based on the color histogram. For each frame, we calculate the L1 distance l between previous color histogram and current one. For all the calculated L1 distances of each video, we calculate their mean μ_l and variance σ_l^2. If $l > \mu_l + 3\sigma_l$, this frame is marked as one candidate for key frames. If the number of candidates for one video is larger than 20, we select top 10 candidates with larger differences as the key frames of this video; otherwise all the candidates are considered as the key frames. The average of extracted key frames for each video is 4 using this algorithm.

For the four-layer hierarchical structure in Foursquare, the annotated venue categories are in the third layer or the second layer. Therefore, our leaf node starts from the third layer. In addition, if there is one venue category of any videos assigned to the second layer, spawn a leaf-node under it and re-assign all the videos from this node of the second layer to this new leaf node [9]. Since we use top-3 layers with additional root nodes (Fig. 3), we define the distribution of parameters in the following forms:

$$\boldsymbol{\beta}_{n,l=3} \sim \mathcal{N}(\boldsymbol{\beta}_{n,l=3}|\boldsymbol{\beta}_{n,l=2}, \lambda_2^{-1}\mathbf{I}) \tag{9}$$

$$\boldsymbol{\beta}_{n,l=2} \sim \mathcal{N}(\boldsymbol{\beta}_{n,l=2}|\boldsymbol{\beta}_{n,l=1}, \lambda_1^{-1}\mathbf{I}) \tag{10}$$

$$\boldsymbol{\beta}_{n,l=1} \sim \mathcal{N}(\boldsymbol{\beta}_{n,l=1}|\boldsymbol{\beta}_{n,l=0}, \lambda_0^{-1}\mathbf{I}) \tag{11}$$

where $\boldsymbol{\beta}_{n,l=3}$ denotes the parameters from the nodes of the third layer (i.e., venue categories of videos) to the final layer of the fusion network. Generally, we set $\boldsymbol{\beta}_{n,l=0} = 0$ and $\lambda_0 = 1$ because of their minimal effect on the remaining parameters [11], [9]. For other hyper-parameters, considering the degree of the effect from different hierarchical layers, we empirically set $\lambda_1 = 5$, $\lambda_2 = 10$ in our experiment.

Similar to [42], we randomly split our dataset into three subsets: 80% of videos are used for training, 10% are used for validation, and the rest 10% are used for testing. All the models are implemented on the Caffe [19] platform.

C. Evaluation Metrics

Similar to [42], we use the standard Macro-F1 and Micro-F1 metrics. Macro-F1 gives equal weight to each class-label in the averaging process; whereas Micro-F1 gives equal weight to all instances in the averaging process. Both Macro-F1 and Micro-F1 reach their best score at 1 and worst one at 0. Let TP_t, FP_t, FN_t denote the true-positives, false-positives and false-negatives for the class-label $t \in T$.

The Macro-F1 and Micro-F1 are defined as follows [10]:

$$\begin{aligned} P_t &= \frac{TP_t}{TP_t + FP_t}, R_t = \frac{TP_t}{TP_t + FN_t} \\ \text{Macro} - \text{F1} &= \frac{1}{|T|} \sum_t \frac{2P_t R_t}{P_t + R_t} \end{aligned} \tag{12}$$

$$\begin{aligned} P &= \frac{\sum_t TP_t}{\sum_t (TP_t + FP_t)}, R = \frac{\sum_t TP_t}{\sum_t (TP_t + FN_t)} \\ \text{Micro} - \text{F1} &= \frac{2PR}{P + R} \end{aligned} \tag{13}$$

Both Macro-F1 and Micro-F1 are informative metrics. The former gives the performance on each category an equal weight in computing the average; the latter gives the performance on each instance an equal weight in computing the average. In fact, Micro-F1 is the accuracy.

D. Evaluation of CPTDL

To demonstrate the effectiveness of CPTDL, we compare our method against the following baselines based on two kinds of VGG16 Networks, namely ImageNet CNN and Places CNN:

- Video-O: Directly using key frames from training videos to fine-tune the ImageNet CNN.
- Image-O: Directly using all the Foursquare images to fine-tune the ImageNet CNN.
- Video-Image-O: first using key frames from training videos, and then using Foursquare images to fine-tune the ImageNet CNN.
- Image-Sel-O: Using the selected Foursquare images to fine-tune the ImageNet CNN.
- O-Late Fusion: Using the selected Foursquare images and key frames from training videos separately to fine-tune two ImageNet CNNs, and then using the max pooling scores as the final prediction.
- Video-S: Directly using key frames from training videos to fine-tune the Places CNN.
- Image-S: Directly using Foursquare images to fine-tune the Places CNN.
- Video-Image-S: First use key frames from training videos, and then using Foursquare images to fine-tune the Places CNN.
- Image-Sel-S: Directly using the selected Foursquare images to fine-tune the Places CNN.
- S-Late Fusion: Using the selected Foursquare images and videos separately to fine-tune two Places CNNs, and then using the max pooling scores as the final prediction.

TABLE I: Performance comparison between our method and the baselines in CPTDL

Method	Macro-F1	Micro-F1
Video-O	13.90%	28.52%
Image-O	12.30%	16.50%
Video-Image-O	13.97%	28.51%
Image-Sel-O	8.30%	12.50%
O-Late Fusion	13.98%	28.72%
CPTDL-O	**15.08%**	**32.50%**
Video-S	15.27%	30.07%
Image-S	13.50%	16.70%
Video-Image-S	15.24%	30.10%
Image-Sel-S	8.50%	12.90%
S-Late Fusion	15.30%	31.05%
CPTDL-S	**15.50%**	**33.60%**

Our method CPTDL-O and CPTDL-S first use key frames from training videos to fine-tuned the network, and then use the fine-tuned model to select foursquare images for further fine-tuning.

The comparative results are summarized in Table I. From Table I, we can observe four key findings: (1) Performance can be improved by taking advantage of Foursquare images for venue category prediction from videos in both two types of networks. Particularly, for ImageNet CNN, compared with Video-O, CPTDL-O achieves a performance improvement of about 1.2 percent in Macro-F1 and 4 percent in Micro-F1, respectively. For Places CNN, there is also the performance improvement of about 0.2 percent in Macro-F1 and 1.1 percent in Micro-F1, respectively. This validates the effectiveness of using images from other platforms to boost the prediction performance of videos in Vine. (2) Our proposed CPTDL performs better than other baselines (e.g., Video-Image-O, O-Late Fusion) that use both images and videos, which validate that our method is effective in learning more discriminative features by taking full advantage of images from Foursquare. (3) We can see that the performance of CPTDL-S is better than CPTDL-O, which indicates that the high-level scene information is more discriminative than the object information for the task of venue category prediction. This is reasonable, since the venue is location-sensitive and scene information is more important than object information. (4) Video-Image-O and Video-Image-S do not perform better than Video-O and Video-S. This is because images from Foursquare are very noisy and may have the semantic drift for a video venue category, which will lead the fine-tuning to the wrong direction. Note that Image-Sel-O and Image-Sel-S perform worse than other methods. The reason is that after image filtering, there are no images for some venue categories and thus the accuracy on these venue categories is 0.

E. Evaluation of Multi-View Feature Learning (MVFL)

After CPTDL, we obtain object and scene features from CPTDL-O and CPTDL-S. Based on the extracted object-scene features, we further verify the effectiveness of MVFL

TABLE II: Performance comparison between our method and the baselines in MVFL

Method	Macro-F1	Micro-F1
O-Fea	15.08%	32.50%
S-Fea	15.50%	33.60%
MVFL-1-4096	15.88%	33.40%
MVFL-2-4096	15.92%	34.20%
MVFL-3-4096	15.89%	33.80%
MVFL-1-8192	16.92%	34.60%
MVFL-2-8192	**17.28%**	**35.20%**
MVFL-3-8192	16.93%	34.70%

TABLE III: Performance comparison between our HCM-FL and other methods on venue category prediction

Method	Macro-F1	Micro-F1
TRUMANN [42]	5.21%	25.27%
DARE [25]	16.66%	31.21%
MVFL-2-8192	17.28%	35.20%
HCM-FL	**18.82%**	**37.40%**

without the venue hierarchical structure prior. We consider the following baselines for comparison:

- Object-Feature (O-Fea). This baseline first extracts the visual features of key frames using the fine-tuned ImageNet CNN from CPTDL, then conduct a mean pooling to obtain the final video representation.
- Scene-Feature (S-Fea). Similar to O-Fea, but use the fine-tuned-Places CNN from CPTDL.

MVFL network fused object features and scene features into a joint feature representation via a multi-layer feed-forward network. We use MVFL-L-D to denote the MVFL network with L fused layers and the units of each fused layer is D. For example, MVFL-1-4096 denotes there is a joint layer with the 4,096 units.

The comparative results are summarized in Table II. We can see that these two kinds of features are complementary. Combining them offers better performance than single features in the task of venue category predication. We further observe that when the number of joint layers is 2 and the units of each joint layer is 8,192, the performance is the best. Therefore, we will use such deep architecture to evaluate our framework.

F. Evaluation of the HCM-FL

Combining CPTDL and MVFL-HVS, we finally verify the effectiveness of our proposed HCM-FL framework. Considering the task of venue category prediction, we choose the following methods as comparison:

- TRUMANN [42]: This is the first work to introduce the task of venue category predication from videos. This method proposed a tree-guided multi-task multi-modal learning model for venue category prediction.
- DARE [25]: This baseline compensated the acoustic modality via harnessing external sound knowledge and developed a deep transfer model for venue category prediction.

Table III shows the experimental results. We can see that after introducing the hierarchical structure prior, the performance has been further improved than MVFL-2-8192. Furthermore, compared with DARE, our HCM-FL has a significant performance gain. HCM-FL can improve the relative performance by 13% for Macro-F1 and 20% for Micro-F1, respectively. This verifies the effectiveness of HCM-FL in jointly utilizing the multi-platform data, multi-view deep features and the hierarchical venue structure prior.

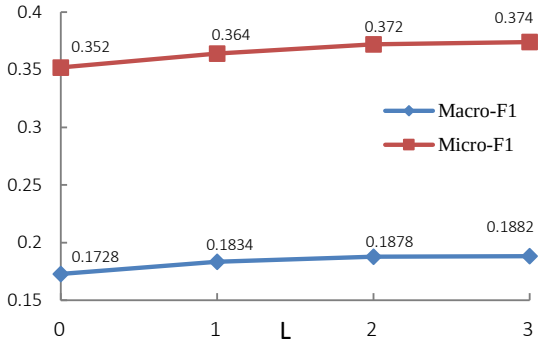

Fig. 7: Comparison of our methods with different hierarchical venue layers.

Figure 7 further compares our framework with different number L of hierarchical layers, where $L = \{0, 1, 2, 3\}$. We can see that there is a consistent increase for both Macro-F1 and Micro-F1 as we increase the layers of venues L from 0 to 3. This shows that representations with higher venue layers become increasingly better at discovering useful features. The reason is that introducing more hierarchical layers makes HCM-FL utilize more prior knowledge to improve the performance. We can also see that the increasing amplitude of the performance becomes smaller with the increase of venue layers. The probable reason is higher venue layers give more minimal effect in improving the prediction performance.

For additional analysis, we also provide venue category-specific results from HCM-FL in Fig. 8. For the space limit, we show the results on 50 venue categories from 188 ones. We report the number that using the following six methods: Video-S, CPTDL-S, Video-O, CPTDL-O, MVFL-2-8192 and HCM-FL. We observe that (1) After CPTDL, there is a consistent performance gain than without transfer deep learning for all the 50 venue categories. The performance of CPTDL-S is better than Video-S, similarly for Video-O and CPTDL-O. This further confirmed that our proposed framework can better leverage the strength of images from other platforms to improve the performance. (2) The scene and object features are complementary. Therefore, the performance of MVFL-2-8192 is better than CPTDL-S and CPTDL-O for many venue categories. There are some failure cases, such as the venue category of Fast Food Restaurant. The reason is that there is a relative large difference for the performance of CPTDL-S and CPTDL-O. Therefore, their fusion leads to an intermediate trade-off between these two features in the prediction performance. Take the venue category Fast Food Restaurant as an example, the performance of CPTDL-O is 1.48% while the performance of CPTDL-S is 47.62%. The performance of MVFL-2-8192 is 12.10%. (3) After introduc-

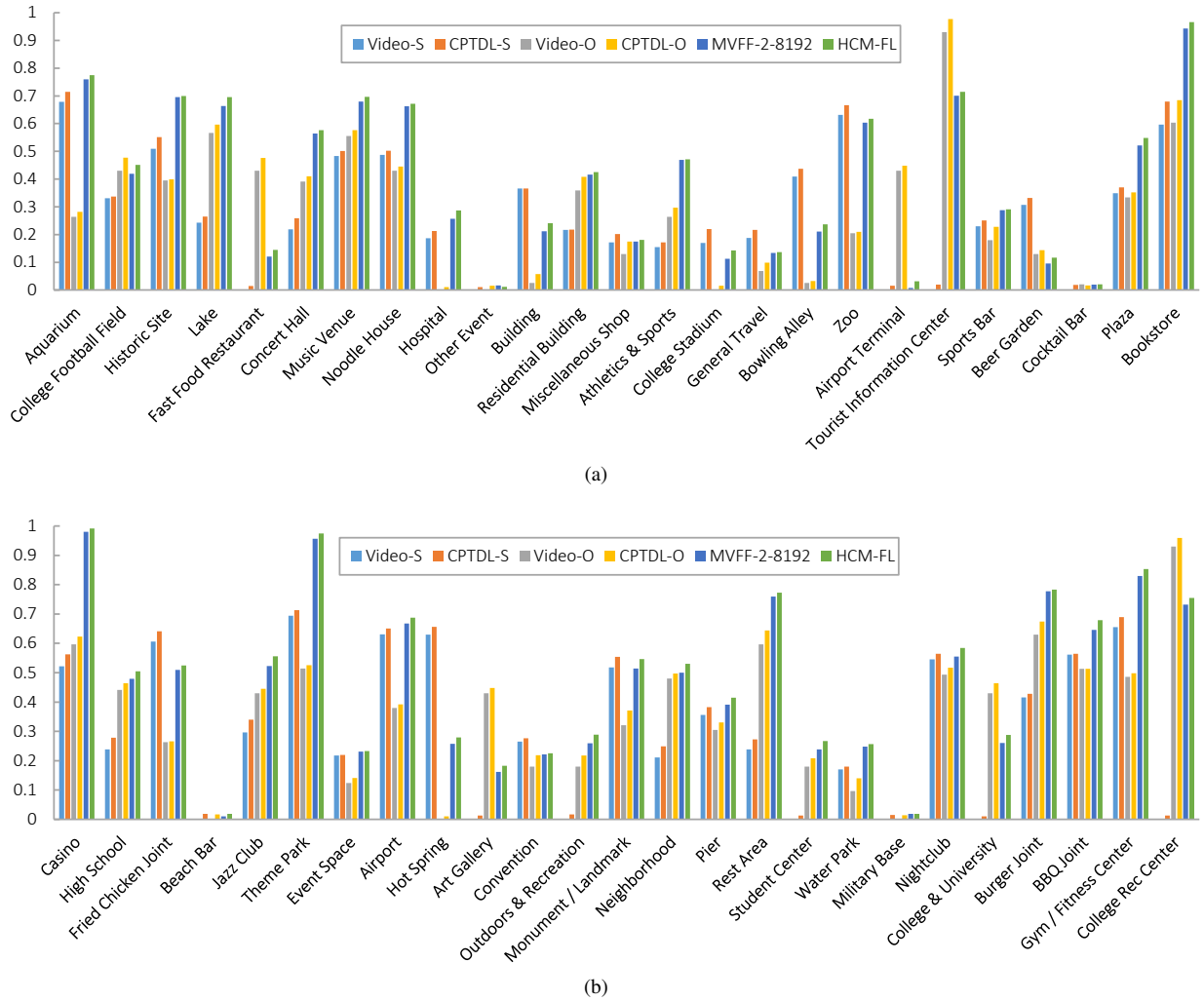

Fig. 8: Per-venue category prediction results for HCM-FL on 50 venue categories

ing the hierarchical venue prior, there is a performance gain for most venue categories out of 50 venue categories.

G. Discussions

Our proposed framework has the relative higher performance gain than existing methods for venue category prediction from videos in Vine. However, there are still relative low prediction performance. This section lays out additional observations that follow from our results to find the probable reasons.

Fig. 9 showed the confusion matrix of HCM-FL over each individual venue category. We can see that our method still does not provide perfect accuracy for some venue categories. We further observe the video data with lower prediction performance and find the following several reasons: (1) Low inter-class variation. For example, as shown in Fig. 10 (a), the videos to describe the Italian Restaurant and Beer Garden venue category are visually similar. Therefore, many test videos from the Italian Restaurant are misclassified into the Beer Garden venue category. In our experiment, the ratio of videos, which are classified to the Beer Garden is about 33.33% while the accuracy for the Italian Restaurant is near 0. (2) High intra-class variation. For example, as shown in Fig. 10 (b), the intra-class variation for the neighborhood venue category is too large. (3) Wrongly-labeled videos. There are also

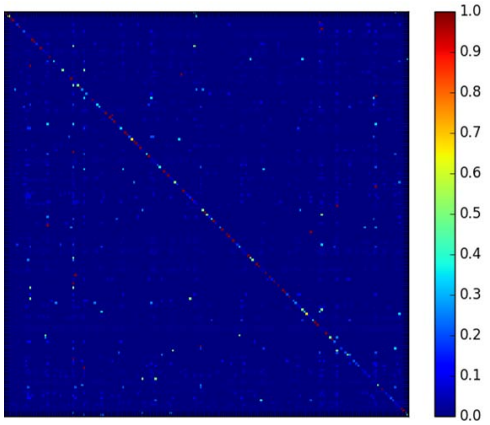

Fig. 9: The detailed comparison over each individual venue category for HCM-FL via the confusion matrix. The row denote true label and the column denote the estimated label. (best viewed under magnification).

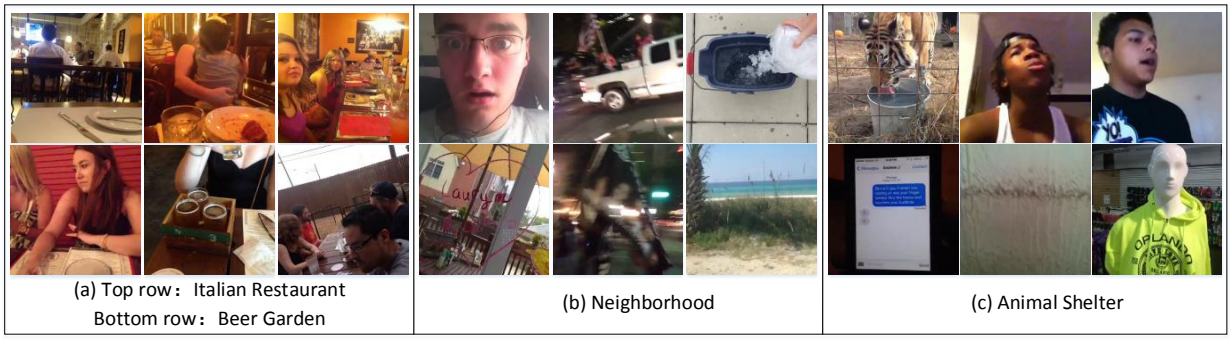

Fig. 10: Some videos from different venue categories to show (a) low inter-class variation, (b) high intra-class variation and (c) some wrong-labeled venue categories

some wrong-labeled videos. Fig. 10 (c) shows some wrongly-labeled videos from the venue category "Animal Shelter". (4) Too few training samples. For example, the number of videos for the Beach Bar is totally only 61. Therefore, its accuracy on Video-O and Video-S is 0. Although after transfer deep learning from Foursquare images, there is an improvement. However, the performance is still low, namely 1.71% and 1.90%, respectively. In order to relieve the above problems, it is probably helpful to combine visual information with textual and other modality information into our framework to solve problem (1) and (2). For problem (3), there are some existing solutions such as [39], [33] to refer and they proposed a solution to train CNNs when there exist mislabeled images in the training set. For problem (4), we can also resort to visual information from more platforms to add more training data.

IV. Related Work

Our work is closely related to the following three research areas: (1) location recognition and prediction, (2) cross-platform analysis and applications, and (3) feature learning with the hierarchial class structure.

A. Location Recognition and Prediction

The goal of location recognition and prediction is to assign the location information to the given text, image or video. It can support a variety of applications, such as event discovery [4], advertising, personalization, location recommendation [5] and location visualization [28]. According to the level of location granularity, there are mainly four types of location prediction, including GPS-level, POI-level, city-level and country-level location prediction.

The task of GPS-level location prediction is to estimate the GPS location given the text [18], images [15], [34] or multi-modal information [32], [3]. For example, Song *et al.* [30] propagated geotags among the web video social relationship graph for video geo-location. In contrast, in many real-world scenarios, especially in social media applications, it is more important for POI-level location prediction. For example, Chen *et al.* [1] mined business-aware visual concepts from social media for recognizing the business venue of images. In addition, there are many retrieval based methods for location estimation [15], [36], [38], [34]. The first work on venue category prediction from videos is from Zhang *et al.* [42]. They proposed a tree-guided multi-task multi-modal learning model for venue category prediction. In addition, they released a large-scale micro-video dataset. Recently, Nie *et al.* [25] fused multi-modal information, especially the acoustic modality for venue category prediction from videos. Among POI-level location prediction, visual landmark recognition and analysis [45], [24], [22], [12] has also been widely studied for its tourism applications. There are also some works on city-level [13] or country-level location prediction [47]. Similar to [42], [25], our work belongs to the POI-level venue category prediction. However, we focus on cross-platform venue category prediction from videos, where the media data from other platforms are effectively exploited to improve the prediction performance.

B. Cross-Platform Analysis and Applications

With the fast development of Web2.0, various media-sharing platforms are gaining more and more popularity with their different types of data and services. Therefore, more and more works resort to cross-platform based study for various applications, such as cross-network based recommendation [23], [40], event detection [2], popularity prediction [26], video recognition and retrieval [8], [14]. For example, Roy *et al.* [26] proposed a novel transfer learning framework that utilizes the knowledge from Twitter to grasp sudden popularity bursts in online content from Youtube. Min *et al.* [23] conducted the recommendation between two platforms, namely photo recommendation from Flickr to Foursquare users and venue recommendation from Foursquare to Flickr users. Gan *et al.* [8] presented a labor-free video concept learning framework by jointly utilizing noisy web videos from Youtbue and images from Google. Different from them, we take advantage of the images from Foursquare for venue category prediction from videos in Vine. In addition, we also jointly utilized multiple semantic cues (e.g., scenes and objects) and the venue structure prior to boost the prediction performance.

C. Feature Learning with the Hierarchical Class Structure

More recently, driven by the great success of Convolutional Neural Networks (CNN) on image recognization tasks [20], [29], a few works attempted to leverage CNN models to learn

feature representations for location recognition. For example, Weyand *et al.* [37] proposed a deep image classification approach in which the world is spatially divided into cells and a deep network is trained to predict the correct cell for a given image. All these works, however, focus on extracting visual or multi-modal features using neural networks, and do not utilize the external class structure knowledge.

Utilizing the hierarchical class structure has become increasingly important for its ability to learn more discriminative features, especially under unbalanced sample distributions over different classes [31]. For example, Gopal *et al.* [11] proposed a Bayesian method to model hierarchical dependencies among class labels using multivariate logistic regression. Different from [11], some works [31], [6] combined the strength of deep neural networks, with tree-based priors, making the deep neural networks work well on unbalanced class distributions. Wang *et al.* [35] exploited hierarchically structured tags from different abstractness of semantics and multiple tag statistical correlations, thus discovered more accurate semantic correlations among different video data, even with highly sparse and incomplete tags. There is a natural geographically hierarchical structure for location information. Some works such as [43], [44] have explored such prior for image location prediction. Zhang *et al.* [37] proposed a tree-guided multi-task multi-modal learning approach to jointly fuse multimodal information, including deep visual features, textual features and audio features from videos for venue category prediction. Different from their work, we use the Bayesian model to exploit the pre-defined hierarchical venue structure prior and combine it with multi-view feature fusion network and transfer deep learning from other platforms to learn more discriminative deep features for venue category prediction.

V. Conclusions

We have proposed a Hierarchy-dependent Cross-platform Multi-view Feature Learning (HCM-FL) framework, which jointly utilized multi-platform data, object-scene deep features and the hierarchical venue structure prior for venue category prediction from videos. In order to utilize the multi-platform data effectively, we proposed a cross-platform transfer deep learning method, which leveraged the complementary nature of the media data from two platforms to reinforce the learned deep network from videos in Vine using the images from Foursquare. In addition, HCM-FL augmented the object-scene feature fusing network with the hierarchical venue structure prior to enable the fusing network to transfer knowledge from related venue categories. Experiments show that HCM-FL achieves better results on two datasets from Vine and Foursqure than existing methods.

As discussed earlier, our work can be extended in the following three directions: (1) We have found that some videos in Vine are wrongly labeled. Therefore, how to improve the robustness of our model under noisy labels is our first direction. Some works [39], [33] proposed a solution to train CNNs when there exist mislabeled images in the training set. (2) We can extend our framework to use data from more platforms. (3) Multi-modal information (e.g., text and audio information) from multiple platforms can also be exploited in the future. In addition, besides the appearance information, temporal information from videos also contains discriminative signals. Thus, we also plan to use the LSTM to capture the sequential features for venue category prediction [21].

References

[1] F. C. B.-C. Chen, Y.-Y. Chen and D. Joshi. Business-aware visual concept discovery from social media for multimodal business venue recognition. In *AAAI*, pages 61–68, 2016.

[2] B. K. Bao, C. Xu, W. Min, and M. S. Hossain. Cross-platform emerging topic detection and elaboration from multimedia streams. *ACM Transactions on Multimedia Computing Communications and Applications*, 11(4):1–21, 2015.

[3] J. Choi and G. Friedland. *Multimodal Location Estimation of Videos and Images*. Springer International Publishing, 2015.

[4] M. Dredze, M. Osborne, and P. Kambadur. Geolocation for twitter: Timing matters. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1064–1069, 2016.

[5] A. Farseev, I. Samborskii, A. Filchenkov, and T. S. Chua. Cross-domain recommendation via clustering on multi-layer graphs. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–204, 2017.

[6] J. Fu, T. Mei, K. Yang, H. Lu, and Y. Rui. Tagging personal photos with transfer deep learning. In *International Conference on World Wide Web*, pages 344–354, 2015.

[7] C. Gan, C. Sun, L. Duan, and B. Gong. *Webly-Supervised Video Recognition by Mutually Voting for Relevant Web Images and Web Video Frames*.

[8] C. Gan, T. Yao, K. Yang, Y. Yang, and T. Mei. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *Computer Vision and Pattern Recognition*, pages 923–932, 2016.

[9] S. Gopal. Large-scale structured learning. In *Thesis*, 2014.

[10] S. Gopal and Y. Yang. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 257–265, 2013.

[11] S. Gopal, Y. Yang, B. Bai, and A. Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. *Advances in Neural Information Processing Systems*, 2012.

[12] T. Guan, Y. Wang, L. Duan, and R. Ji. On-device mobile landmark recognition using binarized descriptor with multifeature fusion. *Acm Transactions on Intelligent Systems and Technology*, 7(1):1–29, 2015.

[13] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49(1):451–500, 2014.

[14] X. Han, B. Singh, V. Morariu, and L. S. Davis. Vrfp: On-the-fly video retrieval using web images and fast fisher vector products. *IEEE Transactions on Multimedia*, PP(99):1–1, 2016.

[15] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[16] X. He, D. Aloi, and J. Li. Portable 3d visual sensor based indoor localization on mobile device. In *The 13th IEEE Annual Consumer Communications Networking Conference*, pages 1125–1128, 2016.

[17] L. Herranz, S. Jiang, and X. Li. Scene recognition with cnns: Objects, scales and dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 571–579, 2016.

[18] M. Hulden, M. Silfverberg, and J. Francom. Kernel density estimation for text-based geolocation. In *AAAI*, pages 145–150, 2015.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[21] M. Liu, L. Nie, M. Wang, and B. Chen. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 970–978, 2017.

[22] W. Min, B. K. Bao, and C. Xu. Multimodal spatio-temporal theme modeling for landmark analysis. *IEEE Multimedia*, 21(3):20–29, 2014.

[23] W. Min, B. K. Bao, C. Xu, and M. S. Hossain. Cross-platform multimodal topic modeling for personalized inter-platform recommendation. *IEEE Transactions on Multimedia*, 17(10):1787–1801, 2015.
[24] W. Min, C. Xu, M. Xu, X. Xiao, and B. K. Bao. Mobile landmark search with 3d models. *IEEE Transactions on Multimedia*, 16(3):623–636, 2014.
[25] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian. Enhancing micro-video understanding by harnessing external sounds. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1192–1200, 2017.
[26] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on Multimedia*, 15(6):1255–1267, 2013.
[27] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010.
[28] J. Sang, Q. Fang, and C. Xu. Exploiting social-mobile information for location visualization. *ACM Transactions on Intelligent Systems and Technology*, 8(3):39:1–39:19, 2017.
[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[30] Y. C. Song, Y. D. Zhang, J. Cao, T. Xia, W. Liu, and J. T. Li. Web video geolocation by geotagged social resources. *IEEE Transactions on Multimedia*, 14(2):456–470, 2012.
[31] N. Srivastava and R. Salakhutdinov. Discriminative transfer learning with tree-based priors. *Advances in Neural Information Processing Systems*, pages 2094–2102, 2013.
[32] M. Trevisiol, J. Delhumeau, and G. Gravier. Retrieving geo-location of videos with a divide and conquer hierarchical multimodal approach. In *ACM International Conference on Multimedia Retrieval*, pages 1–8, 2013.
[33] A. Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *arXiv:1706.00038*.
[34] N. Vo, N. Jacobs, and J. Hays. Revisiting im2gps in the deep learning era. In *IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.
[35] J. Wang, X. Zhu, and S. Gong. Video semantic clustering with sparse and incomplete tags. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 3618–3624, 2016.
[36] S. Wang and S. Jiang. INSTRE: A new benchmark for instance-level object retrieval and recognition. *TOMCCAP*, 11(3):37:1–37:21, 2015.
[37] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. 2016.
[38] T. Weyand, I. Kostrikov, and J. Philbin. *PlaNet - Photo Geolocation with Convolutional Neural Networks*. Springer International Publishing, 2016.
[39] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. pages 2691–2699, 2015.
[40] M. Yan, J. Sang, C. Xu, and M. S. Hossain. Youtube video promotion by cross-network association: @britney to advertise gangnam style. *IEEE Transactions on Multimedia*, 17(8):1248–1261, 2015.
[41] J. Zahalka, S. Rudinac, and M. Worring. Interactive multimodal learning for venue recommendation. *IEEE Transactions on Multimedia*, 17(12):2235–2244, 2015.
[42] J. Zhang, L. Nie, X. Wang, X. He, X. Huang, and T. S. Chua. Shorter-is-better: Venue category estimation from micro-video. In *ACM on Multimedia Conference*, pages 1415–1424, 2016.
[43] X. Zhang, X. Hu, and Z. Li. Learning geographical hierarchy features for social image location prediction. In *International Conference on Artificial Intelligence*, pages 2401–2407, 2015.
[44] X. Zhang, X. Hu, S. Wang, Y. Yang, Z. Li, and J. Zhou. Learning geographical hierarchy features via a compositional model. *IEEE Transactions on Multimedia*, 18(9):1855–1868, 2016.
[45] Y. T. Zheng, M. Zhao, Y. Song, and H. Adam. Tour the world: Building a web-scale landmark recognition engine. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1085–1092, 2009.
[46] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems*, 1:487–495, 2014.
[47] A. Zubiaga, A. Voss, R. Procter, M. Liakata, B. Wang, and A. Tsakalidis. Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, PP(99):1–1, 2016.