HOW SCENES IMPLY ACTIONS IN REALISTIC VIDEOS?

Hongsong Wang, Wei Wang, Liang Wang

Center for Research on Intelligent Perception and Computing, Institute of Automation
National Laboratory of Pattern Recognition, Chinese Academy of Sciences
{hongsong.wang, wangwei, wangliang}@nlpr.ia.ac.cn

ABSTRACT

People drive on the road and eat in the kitchen. Can the road imply driving or the kitchen imply eating? This paper addresses such a problem by studying the relations between actions and scenes. To get effective scene representation, we use a deep convolutional neural networks (CNN) model trained from a scene-centric database to predict scene responses for videos. We employ two encoding schemes based on frame features to represent the scene and its changes, respectively. We conduct experiments on two challenging datasets, HMDB51 and Hollywood2, and compare action recognition results of different encodings based on different scene features. Our results demonstrate that scene features, when combined with motion features, improve the state-of-the-art results for action recognition. Finally, we explore the relationship between actions and scenes by analyzing scene preferences to a particular action qualitatively and quantitatively.

Index Terms— Action recognition, scene features, motion features, static scene encoding, dynamic scene encoding, scene preferences of action

1. INTRODUCTION

Action recognition has been an active research area due to many applications. The attention of research has shifted from simple actions in controlled environments [1, 2] to complex actions in realistic scenarios [3, 4, 5]. Most approaches place emphasis on designing handcrafted descriptors to represent motions. For example, the very popular dense trajectories [6, 7] extract hand-crafted features (e.g., TRJ, HOG, HOF, MBH) based on trajectories of interest points.

Besides movements, scene and object cues are also important for action recognition. For example, Wu et al. [8] detect and analyze the sequence of objects manipulated by a user. Han et al. [9] design contextual scene descriptors to encode structural relations between object parts. Ullah et al. [10] use non-local cues defined by several techniques including object detection and foreground segmentation to improve action recognition. Ikizler-Cinbis et al. [11] integrate multiple feature channels from several entities such as objects and humans to identify actions. These methods either use a few explicit object detectors, or treat moving regions as object candidates. And they simply regard the scene as background or contextual region around object. Thus, they cannot reveal the relations between actions and scenes (objects).

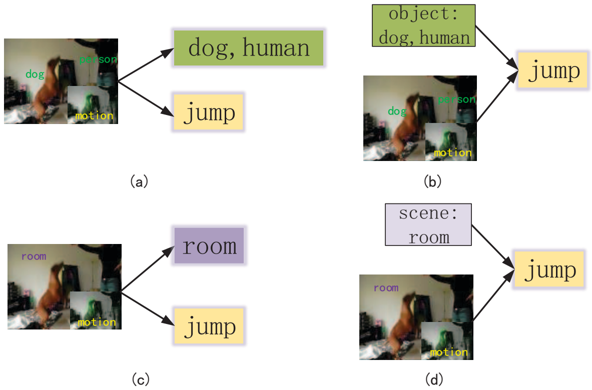


Fig. 1. Four action recognition cases considering object or scene. (a) Joint objects and action recognition [12]. (b) Use object encodings as additional features [13]. (c) Joint scene and action recognition [3]. (d) Use scene encodings as additional features. We study this problem since it has not yet been explored. (Best viewed in color.)

There are also some studies focusing on discovering relations between actions and scenes (objects). For example, Jain et al. [13] conduct an empirical study of encoding 15,000 objects categories for actions, and reveal that object responses improve action classification results. Xu et al. [12] collect a dataset with labels of both objects and actions, and demonstrate that joint inference over objects and actions outperforms independent inference over them. Marszalek et al. [3] develop a joint framework for action and scene recognition. However, their method relies on movie scripts for training and cannot apply for general videos. Hence it still remains an open problem to study the encoding of scenes for action recognition. We visualize and compare the four cases in Fig. 1.

Inspired by recent progress of deep convolutional neural networks (CNN) in large-scale scene recognition [14, 15], we decide to use a CNN model trained from large scene-centric databases (e.g., Places [14]) to predict scene responses for videos. The proposed framework is shown in Fig. 2. First,

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

different layers of scene features, *scene category* (prob) and *scene attribute* (fc6) are extracted for each frame using [16]. Then, we use two encoding schemes, *static scene encoding* and *dynamic scene encoding*, to describe the scene and its changes, respectively. Finally, these features are combined with motion features (e.g., [7]) for action recognition.

Our experiments demonstrate that scene features improve the state-of-the-art results for action recognition, better than object features [13] for action recognition. We also reveal that for *scene category*, *static scene encoding* is better than *dynamic scene encoding*, and the opposite conclusion is obtained for *scene attribute*. Finally, we analyze scene preferences of a particular action qualitatively and quantitatively.

Our paper makes the first study of encoding different scene categories for action recognition. We employ two encoding schemes to model static and dynamic scenes, respectively, and our results show that scene features do improve the state-of-the-art results for action recognition.

2. METHOD

In this section, we aim to discover relations between action and scene. We want to show how scenes imply actions in realistic videos. The pipeline of action classification is illustrated in Fig. 2. Since many previous action recognition works focus on motion, there are a variety of choices of motion features. For example, we can use Fisher vector encoding of local descriptors of the improved dense trajectories [7]. We employ two different encoding schemes based on scene features of different frames. The details are shown as follows.

2.1. Deep Scene Feature

As deep convolutional neural networks have achieved great success for scene classification, we use the in-house models to extract scene features for video frames. Places [14] is the first large-scale scene classification database with 205 scene categories and with over 7 million labeled pictures. As 205 scene categories cover most background of actions, we do not use a larger database such as Place2 [15]. We adopt the Places205-VGGNet model [16] due to its state-of-the-art performance. This deep convolutional architecture has 16 layers with small sizes of convolutional kernel (3 × 3), stride (1 × 1), and pooling window (2 × 2), following original implementation of [17]. To analyze differences of features from different layers, we extract category level features (prob) and attribute level features (fc6), respectively.

2.2. Static Scene Encoding

Due to temporal continuity over time, scene features of adjacent frames are quite similar. To get scene representations for videos, we can aggregate the statistics of these features over frames. We simply use *max pooling* since the performance is slightly better than average pooling. This operation captures the primary scene features, and we call this *static scene encoding*.

2.3. Dynamic Scene Encoding

Some actions have scene changes, and accordingly videos have shot transitions (e.g., abrupt transitions). The change of scene also reflects the meaning of an action. For example, diving can be characterized by the scene changes from platform (air) to water. We use *rank pooling* [18] to model scene evolutions for actions. It first regards the frame sequence as an ordered list. Then, learns a ranking function based on frame features to keep the order constraints. Finally, the parameters of this ranking function are used as video representation. For more details, please refer to [18]. This operation captures the scene evolution, and we call this *dynamic scene encoding*.

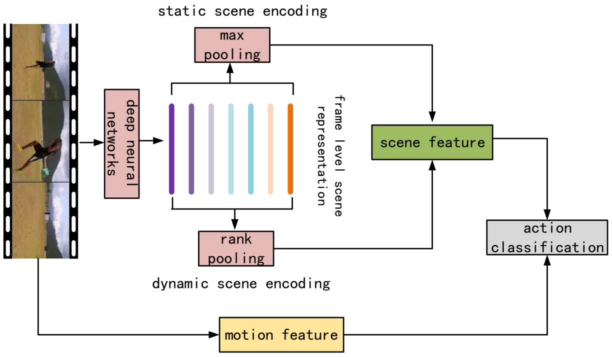


Fig. 2. The pipeline of scene based action recognition. First, a CNN model is used to extract scene features for each frame. Then, *max pooling* and *rank pooling* [16] are used to encode features of the scene and its changes, respectively. We use these two encodings as additional scene features for action recognition.

3. EXPERIMENT AND ANALYSIS

Now we present detailed experimental results and discussions. For multi-class classification, we use libsvm [19] with linear kernel ($C = 100$), and apply *one-against-rest* strategy, following [7].

3.1. Datasets

For action recognition, we conduct experiments on two challenging datasets, namely HMDB51 [4] and Hollywood2 [9].

HMDB51. This dataset is a large collection of realistic videos from various sources, including movies and web videos. It contains 6,766 video clips from 51 action classes, each having at least 101 videos. We follow the standard evaluation scheme of [4] using three train/test splits. We use the original set, which is not stabilized, and report the average classification accuracy over these three splits.

Hollywood2. This dataset is collected from 69 different Hollywood movies. It has 12 action classes. Following the evaluation protocol of [9], we use the clean training subset of 823 videos and the clean test subset of 884 videos. Training and test videos are selected from different movies. The mean average precision (mAP) over all classes is used to measure the final performance.

3.2. Scene Implies Action

In Table 1, we provide action recognition results of *static scene encoding* and *dynamic scene encoding* for *scene category* and *scene attribute*, respectively. We also show feature level fusion results by concatenating these two encodings.

We observe that for *scene category*, *static scene encoding* is better than *dynamic scene encoding*. And the opposite conclusion is reached for *scene attribute*. Another observation is that we can get the best performance by combining the two encoding methods. As *static scene encoding* captures the primary scene features, and *dynamic scene encoding* models the scene evolution, they are both effective, and complementary with each other.

3.3. Scene Beats Object

In [13], Jain et al. show that object encodings can be used for action recognition. Here we compare our scene encodings with their object encodings. For scene features, we use *scene attribute* due to the better performance. For motion features, we use [20] for the HMDB51, and [7] for the Hollywood2, in order to be consistent with [13]. Similar to [13], we concatenate scene features with motion features for action recognition. The results are shown in Table 2.

We can see that on the Hollywood2 dataset, scene encodings (50.3%) perform much better than object encodings (38.4%). And on the HMDB51 dataset, scene encodings are slightly better than object encodings.

For both datasets, we can see that the combination of scene and motion is much better than that of object and motion. For example, our method outperforms [13] by 3.4% on the Hollywood2 dataset. The results indicate that scene encodings are better than object encodings for action recognition.

3.4. Analysis of Particular Action

To analyze the recognition difference of different actions, we show the result for each action. We only give the results for the Hollywood2 due to space limitations, see Fig. 3.

We can see that for some actions (e.g., *DriveCar*, *Eat*, *Run*), the recognition results based on scene cues are very promising. For example, the average precision of *DriveCar* based on *scene attribute* is 77.5%, much higher than that of other actions. This is not difficult to understand because people drive on the road and eat in the kitchen. A particular scene goes hand in hand with these actions. On the other hand, for some other actions (e.g., *HandShake*, *HugPerson*), the recognition results based on scenes are not good.

We can also see that scene and object can be used as additional cues to help action recognition. Nearly for all actions, both object encodings and scene encodings can improve action recognition results.

Table 1. Action recognition results of *static scene encoding* and *dynamic scene encoding* for both *scene category* (prob) and *scene attribute* (fc6).

(%)	*scene category*			*scene attribute*		
	static	dynamic	fusion	static	dynamic	fusion
Hollywood2	28.0	20.6	30.8	40.3	43.0	50.3
HMDB51	22.7	14.1	23.8	34.9	35.3	39.0

Table 2. Comparison of scene encodings with object encodings for action recognition.

(%)	motion	object [13]	scene	object + motion [13]	scene + motion
Hollywood2	64.6	38.4	50.3	66.4	**69.7**
HMDB51	66.8	38.9	39.0	71.3	**73.6**

3.5. State of the Art Performance

In Table 3, we compare our results with the best methods in the literature on these two datasets.

On the HMDB51 dataset, our method achieves the state-of-the-art performance, outperforming the existing best result [13] by 2.3%.

On the Hollywood2 dataset, our method also outperforms many recent state-of-the-art results, but is inferior than [18], which presents a new encoding method for motion by modeling appearance evolution. We believe we can further improve the result by using [18] as motion features on this dataset.

3.6. Scene Preferences of Action

In the previous subsections, we have demonstrated that scene can be used to improve the state-of-the-art action recognition results. In this subsection, we explore the relation between action and scene by analyzing scene preferences of a particular action.

Qualitative analysis. We first provide qualitative analysis of relations between actions and scenes. For an action dataset, we sum up the category level features (prob) of videos for each action class in the training set, and treat the results as the contribution of scene categories to action classes. We sort the *scene category* responses in a descending order and select the top 100 scene categories for each action class. As we do not know the specific names of *scene attribute* in our network, we do not conduct similar analysis for attributes (fc6).

The visualizing results of four actions in UCF101 are shown in Fig. 4. We can see that scene responses are semantically related to the action classes. For example, for action *Bowling*, the scene *bowling-alley* is most obvious. And for action *Rafting*, the scene *raft* and *iceberg* are more obvious than others.

Quantitative results. In the above qualitative analysis, we find that most actions have their preferred or high scene responses. Here we assess how these preferred scenes contribute to action recognition. For each sample, we select the

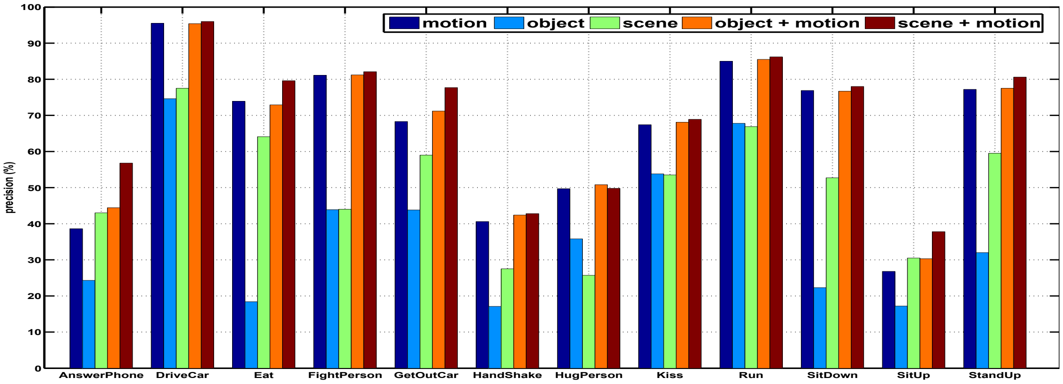
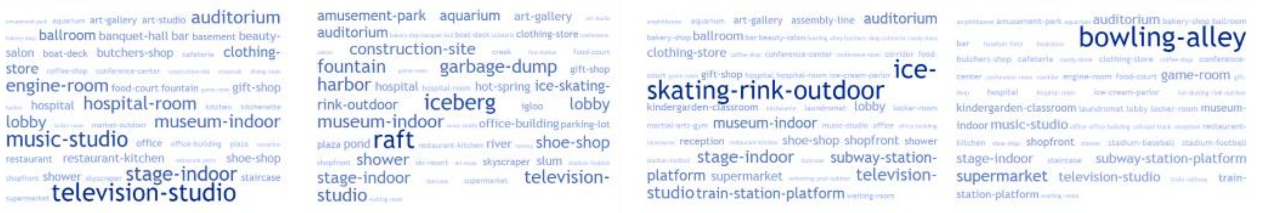


Fig. 3. Action recognition result for each action based on scene features on the Hollywood2 dataset.


art-gallery art-studio auditorium
ballroom banquet-hall bar basement beauty-salon boat-deck butchers-shop clothing-store
engine-room food-court fountain gift-shop
hospital hospital-room
lobby museum-indoor
music-studio office
restaurant restaurant-kitchen shoe-shop
shower stage-indoor staircase
television-studio
amusement-park aquarium art-gallery
auditorium clothing-store
construction-site
fountain garbage-dump gift-shop
harbor hospital hot-spring ice-skating-rink-outdoor iceberg igloo lobby
museum-indoor office-building parking-lot
plaza pond raft river shoe-shop
shower skyscraper slum
stage-indoor television-studio
art-gallery assembly-line auditorium
ballroom
clothing-store conference-center
gift-shop ice-skating-rink-outdoor
kindergarden-classroom lobby
museum-indoor
reception shoe-shop shopfront shower
stage-indoor subway-station-platform supermarket television-studio train-station-platform
amusement-park auditorium
bowling-alley
bar
butchers-shop cafeteria clothing-store conference-center
game-room
hospital
kindergarden-classroom museum-indoor music-studio
kitchen shopfront
stage-indoor subway-station-platform
supermarket television-studio train-station-platform

(a) Drumming

(b) Rafting

(c) IceDancing

(d) Bowling

Fig. 4. Qualitative results of visualizing contributions of scene categories for four actions from the UCF101 dataset.

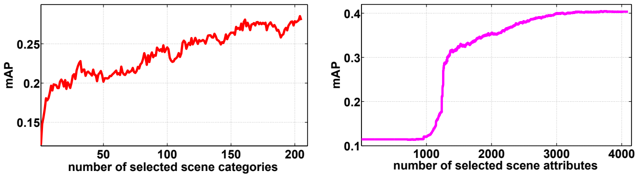
Table 3. Comparison of our method with the state-of-the-art methods in the literature.

(%)	HMDB51	Hollywood2
Zhu et al. [21]	54.0	61.4
Oneata et al. [22]	54.8	63.3
Wang et al. [7]	57.2	64.3
Peng et al. [20]	66.8	–
Lan et al. [23]	65.1	68.0
Ni et al. [24]	65.5	66.7
Fernando et al. [18]	63.7	**73.7**
Jain et al. [13]	71.3	66.4
Our method	**73.6**	69.7

top K most responsive scene categories, and only use these for action recognition. We evaluate this impact on action recognition by varying the value of K. Similar experiment is performed for scene attributes. For simplicity, we only show the results of the Hollywood2 dataset in Fig. 5.

The conclusions are as follows. For *scene category*, we only need about 150 categories to obtain the optimal result. And for *scene attribute*, 3,000 attributes are sufficient to reach the maximum value. Specifically, using about 50% of the total attributes, we can achieve about 90% of the best performance.

One of the drawbacks of this analysis is that the responsive scene components might not be discriminative for action recognition. To conduct precise analysis, an additional classifier should be used to predict the probability relevant with a specific action for each scene component. This improved scheme requires a lot of work, so we consider it as a future plan.



(a) category

(b) attribute

Fig. 5. Recognition results of selected top K most responsive values of (a) *scene category* and (b) *scene attribute* on the Hollywood2 dataset.

4. CONCLUSIONS

In this paper, we study the relations between actions and scenes, and leverage encodings of scenes for action recognition. We use a CNN model trained from a large scene-centric database to predict scene features for video frames. Then, different encoding methods (*static scene encoding* and *dynamic scene encoding*) are adopted to get effective scene representations. Our experiments show that scene features can improve the state-of-the-art results of action recognition on the two challenging datasets (HMDB51 and Hollywood2).

5. ACKNOWLEDGEMENT

This work was supported by the National Basic Research Program of China (2012CB316300) and National Natural Science Foundation of China (61175003, 61135002, 61202328, 61420106015, U1435221).

6. REFERENCES

[1] Christian Schüldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *ICPR*. IEEE, 2004.

[2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," in *ICCV*. IEEE, 2005.

[3] Michael Marszalek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *CVPR*. IEEE, 2009.

[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*. IEEE, 2011.

[5] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CRCV-TR-12-01*, 2012.

[6] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011.

[7] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *ICCV*. IEEE, 2013.

[8] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M Rehg, "A scalable approach to activity recognition based on object use," in *ICCV*. IEEE, 2007.

[9] Dong Han, Liefeng Bo, and Cristian Sminchisescu, "S-election and context for action recognition," in *ICCV*. IEEE, 2009.

[10] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev, "Improving bag-of-features action recognition with non-local cues.," in *BMVC*. Citeseer, 2010.

[11] Nazli Ikizler-Cinbis and Stan Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *ECCV*. 2010, Springer.

[12] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso, "Can humans fly? action understanding with multiple classes of actors," in *CVPR*. IEEE, 2015.

[13] Mihir Jain, Jan C van Gemert, and Cees GM Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?," in *CVPR*. IEEE, 2015.

[14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *NIPS*. MIT Press, 2014.

[15] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva, "Places2: A large-scale database for scene understanding," *arXiv preprint arXiv:1505.01130*, 2015.

[16] Limin Wang, Sheng Guo, Weilin Huang, and Yu Qiao, "Places205-vggnet models for scene recognition," *arXiv preprint arXiv:1508.01667*, 2015.

[17] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars, "Modeling video evolution for action recognition," in *CVPR*. IEEE, 2015.

[19] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *TIST*, 2011.

[20] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng, "Action recognition with stacked fisher vectors," in *ECCV*. Springer, 2014.

[21] Jun Zhu, Baoyuan Wang, Xiaokang Yang, Wenjun Zhang, and Zhuowen Tu, "Action recognition with actons," in *ICCV*. IEEE, 2013.

[22] Dan Oneata, Jakob Verbeek, and Cordelia Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *ICCV*. IEEE, 2013.

[23] Zhenzhong Lan, Ming Lin, Xuanchong Li, Alexander G Hauptmann, and Bhiksha Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *CVPR*. IEEE, 2015.

[24] Bingbing Ni, Pierre Moulin, Xiaokang Yang, and Shuicheng Yan, "Motion part regularization: Improving action recognition via trajectory group selection," in *CVPR*. IEEE, 2015.