1 Team details

- Team Name: Northumbria University Newcastle
- Leader: Yi Zhou
- Address: PB238, Pandon Building, Northumbria University, Newcastle upon Tyne, United Kingdom. NE1 8ST

  Email: yi2.zhou@northumbria.ac.uk
- Other members: Li Liu; Daniel Organisciak; Ling Shao
- Team website URL (if any)
- Affiliation: Northumbria University Newcastle

2 Contribution details

- Method Title: Coarse-to-Fine CNN exploiting Object and Scene Cues for Recognizing Culture Events
- Final score (Not released)
- General Description: Our method is mainly based on the widely used Convolutional Neural Networks. We separately extract object and scene information from two different models. In addition, we also adopt the multi-resolution and local patch selection techniques to help train the models. The architecture CaffeNet of pre-trained model learned on ImageNet and Places205 dataset is used, and we fine-tune the models using training and validation dataset of culture event images under our proposed architecture.
- References

[1]Wang, Limin, et al. ”CUHK SIAT submission for THUMOS15 action recognition challenge.” THUMOS15 Action Recognition Challenge 3.4 (2015).

[2]Salvador, Amaia, et al. ”Cultural Event Recognition with Visual ConvNets and Temporal Models.” arXiv preprint arXiv:1504.06567 (2015).

[3]S. Park and N. Kwak, ”Culture Event Recognition by Subregion Classification with Convolutional Neural Network”.

[4]Zitnick, C. Lawrence, and Piotr Dollr. ”Edge boxes: Locating object proposals from edges.” Computer VisionECCV 2014. Springer International Publishing, 2014. 391-405.

[5]Jia, Yangqing, et al. ”Caffe: Convolutional architecture for fast feature embedding.” Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

- Representative image / diagram of the method see in Figure 1.

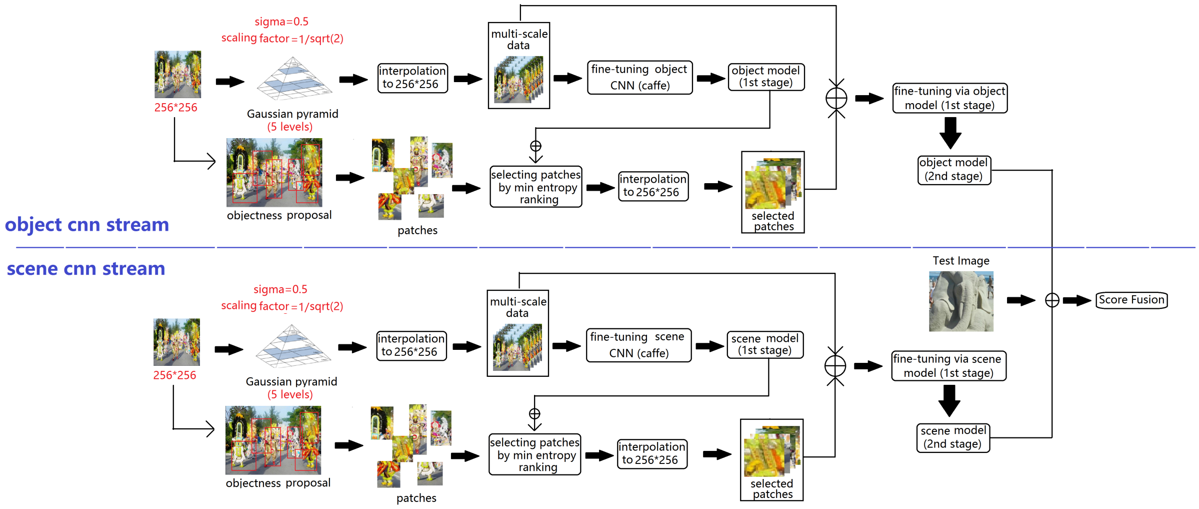

Figure 1: Working flow of the proposed method.

3 Data Preprocessing

- Features: We adopt fully connected layer features from our fine-tuned CNN models consisting of original global images and selected local salient patches. Object-based CNN and Scene-based CNN are both explored.

- No Dimensionality reduction technique

- No Segmentation strategy

- Multi-resolution technique

  All the training and validation data are first used to construct 5 level Gaussian Pyramid as Figure 1. Then all the data from each level are interpolated to 256 × 256 for training the coarse-level CNN model. This first stage net is obtained by fine-tuning the pre-trained model of Imagenet and Places, respectively.

- Local Patch Selection

  After training the first stage model on all the whole images, we then use Edge Boxes[4] to locate object proposals in each image. In particular, we reject the proposals which are small and too tall/fat regions. In this way, we more focus on those big local regions containing rich information of corresponded culture event. Then we classify all the remained proposals using the first stage trained models and discard incorrectly classified patches. We further select those correctly classified patches with top 20% low entropy for each event category based on minimum entropy ranking.

Finally, we combine the selected patches (interpolated to 256×256) with the multi-scaled data (interpolated to 256×256) together to fine-tune the first stage net and then obtain the second stage model which is beneficial with both global and local information.

- Combination of Object and Scene Nets

  All above procedure are applied on both Object CNN stream and Scene CNN stream. Finally, the scores are mean-fused from the two nets.

4 Classification details

- Classifier: CNN with the softmax layer to do the classification
- Large scale strategy (if any)
- Compositional model used (scene context representation), i.e. pictorial structure (if any)
- Other technique/strategy used not included in previous items FOR CLASSIFICATION (if any)

5 Global Method Description

- Total method complexity analysis: The training stage of our method is complex and very time-consuming, but test stage is simple.
- Pre-trained CaffeNet based on ImageNet and Places205 dataset are used to fine-tune our models.
- Qualitative advantages of the proposed solution

  Multi-resolution technique: Since all the train, validation and test data are collected from the Internet, resolution of images are various significantly. So the multi-resolution technique is helpful for training the model to learn different culture event features under different resolutions. It can improve the classification accuracy at some extent.

  Local patches selection: Some salient subregions containing key objects or local activities are always the significant cues for discriminate different culture events. So we further fine-tune our models to learn local features in the second training stage.

  Combination of Object CNN stream and Scene CNN stream: The object net aggregates important information for recognizing event from the perspective of object, while the scene net performs event recognition with the help of scene context. So the combination of the two nets is beneficial for the results.
- Results of the comparison to other approaches (if any)

- Novelty degree:

  (1) Multi-resolution technique is used in our training phase for culture event recognition tasks.

  (2) We also combine local patches selection and Multi-resolution data together to construct a coarse-to-fine CNN learning architecture.

  (3) We propose two CNN learning streams beneficial from both object and scene and then fuse them together to produce a final decision.

6 Other details

- implementation platform: Caffe[5] on Ubuntu; CPU: 4.4GHz; RAM: 32G-B; GPU: NVIDIA TITAN X
- Human effort: preprocessing the training and validation data; programming based on our algorithms; tuning the parameters for training to achieve the best results.
- Training/testing expended time: one month.
- General comments and impressions of the challenge: The competition helps us to have a deep understand about CNN, and gives us some inspiration to extend CNN on other interesting applications.