Fan Gongxiu Honors College

Advertisement Recommendation System Based on Improved Saliency Detection and Action Recognition Algorithm

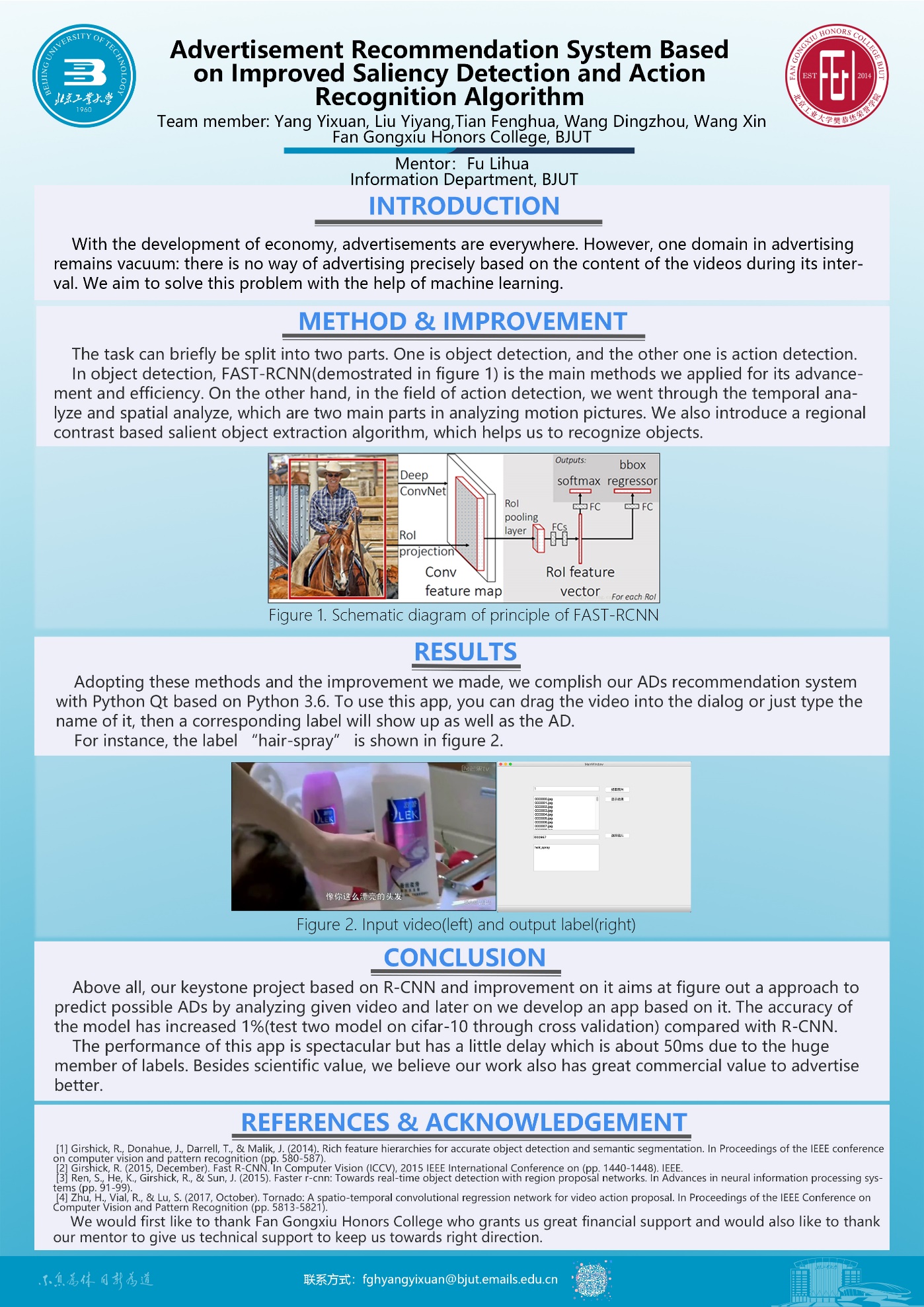
A Keystone Project

Final Report

Authors: Yang Yixuan\*, Liu Yiyang, Tian Fenghua, Wang Dingzhou, Wang Xin

Supervised by: Fu Lihua, the Information Department of Beijing Univers-ity of Technology

Date: December 23, 2018

**

# Executive Summary

Over the last whole year, we have been devoting time on this project, which finally received a good result as a reward to our devastating hard word, and which, along with its innovation, is the most important reason why our revolutionized work was done by our five group members. We decided to work on this project because of the willingness and desire of changing the poor efficient way where commercials are publicized, and improving the experience of users where they are exposed to commercials. With innovation, devotion and responsibility of changing the world, we finally made it.

In this project, several significant milestones with shiny functions were realized with our simple but wonderful lines of code, including:

* Accurately determination on significance of objects appearing in videos.
* Object recognition in a continuous video.
* Advertisement recommendation based on significant object appearing in the video.
* The first realize on combination of digital video processing techniques and commercial activities.

These achievements are just meeting with the goals preliminarily set at the very beginning of the project.

Our achievement is revolutionary not just because of its comprehensive utilization of avant-garde techniques, but its further deep impact on the way where commercials are circulated and functioned. However, there is still going to be multiple issues to be achieved in the future to make our work being promoted and transmuted in the future so that our work can really improve users’ experience. What we have achieved now needs to be promoted both technically and practically. Next, we are looking for advertisement providers and video service provider, asking for further cooperation with market and our techniques. Due to the limitation of the reimbursement policy, our team has not yet had a computer equipped with a high-performance GPU for the project. The model we built only runs on our personal computer. However, the performance of personal computers is low, which limited the size of our model.

# Background

With the development of economy, advertisements are everywhere. Advertising used to be “blind”, since all the users were given the same advertisements. It is obvious that such advertisements do not have high efficiency. Later on, by utilizing the Internet to gather personal data, advertisements finally get to be deployed “precisely” according to user browsing history, habits, and so on.

However, one domain in advertising remains vacuum: there is no way of advertising precisely based on the content of the videos during its interval. Not anymore. With the help of machine learning, we aim to solve this problem. We analyze the content of the video, using which to help corporations to advertise more precisely. Besides scientific value, we believe our work also has great commercial value to advertise better.

People are extremely annoyed when commercials fly in front of their eyes. Often it usually has nothing to do with their needs. What we are thinking is to take full advantage of the information left by Internet users and analyze what commercials to show. However, we also have to face the fact that a large quantity of information is in the form of images and videos, which are impossible to be used directly. Our job is to extract useful information from the videos and take use of them. We hope to adjust algorithms we have at hand and apply it to commercial circulation. In such situations, commercials will be shown intentionally for better accuracy. People will see commercials about what they really need and will not see irrelevant commercials.

Extracting information from videos requires obtaining each movement, which is called action detection. It tells when a movement occurs, and with the help of action recognition, we can know what happens in that specific moment. However, preliminary work such as object detection and recognition have to be done. Therefore, we can divide the task into several mini-tasks.

Object detection and action detection of a video is fundamental to information extraction, and deep learning-based algorithms have provided us with a way to achieve such a goal and have a good improvement of recognition. It aims to give the exact time and location of an event. Our first step concerns object detection.

There have been many pioneers focusing on improving object detection methods. In year 2014, Girshick put forward Region Convolutional Network (R-CNN) to detect objects of a single picture. It preliminarily selects several regional proposals, extract its features using CNN, and send these features to Super Support Vector Machine (SVM) to judge whether it belongs to any certain categories. He improved it with a new method, Fast R-CNN by directly sending the whole picture by normalization without regional proposals. It avoided repeated calculation and save training time. It also unified category classification and location classification into a single deep network. It requires less space than R-CNN and save storing space. In the same year, new promotion has been made that the total analyzing rate can be as fast as 17 fps, the Faster R-CNN. It replaced the selective search methods in Fast R-CNN with regional production network. Moreover, in 2017, K. He put forward Mask R-CNN [4] for a faster performance.

# Objectives

The Project is mainly achieved by our five members at school. The tasks can be divided into preliminary research and realization. During preliminary research, we completely learnt the sift algorithm and HOG processing, and further we understand more about computer vision, convolutional neural network and saliency detection. During this period, the R-CNN processing method was learned, and the image recognition and classification can be achieved with high accuracy. After that, we understand and reproduce the method used to process the video called T-CNN algorithm. We have a certain understanding and basis for processing video.

Our main goal is to make an Advertising Recommendation System, which could extract the significance area of each frame of the image in videos and make object recognition and action recognition. Paper review can take a large part.

Searching for the newest algorithm and learn it, after that we aim to prove it. We want to improve the algorithm, so we must know how it works and what the defects of these methods), include RC, HC and mask R-CNN.

We introduce a regional contrast based salient object extraction algorithm, which simultaneously evaluates global contrast differences and spatial weighted coherence scores. These saliency maps are further used to initialize a novel iterative version of GrabCut for high quality salient object segmentation.

For the part of object recognition, we made a detection of the image, first find out the ROI (Region Of Interest) in images, for each ROI area using “ROIAlign” make pixel correction. Then for each ROI use the design of Fast-CN classification framework to predict different instance belongs to right or wrong, and get the segmentation result of the image instance.

Later, use the mask encodes the spatial layout of the input target. And use the m\*m matrix to predict each ROI rather than using a vector, this ensures that the spatial information in the ROI will not be lost.

First, we use the Histogram Contrast algorithm (HC), which based on the color statistical characteristics of the input image and detect it.

Second, we use Region Contrast algorithm (RC) to combine spatial relations and regional level contrast together, the picture first divided into several regions, with each region and their regional contrast weighting and to calculate the regional significance.

Try the latest motion recognition algorithms, such as behavior recognition based on unsupervised learning and behavior recognition based on convolutional neural networks. Use the program to reproduce the algorithm, adjust the code so that the program operation results are good, and stable.

# Methodology

Our work is mainly in the laboratory room, and has written code for the saliency detection and deep residual network. After obtaining a large number of training sets and test sets on ImageNet, it was completed by a Titan X GPU for two days. The accuracy rate can reach more than 95%.

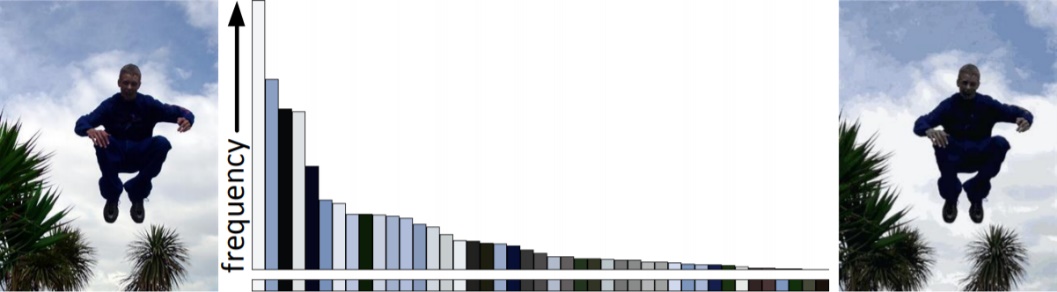
Since the input file is in video format, we learn to frame the video first. After that, the overall idea of the processing after our video frame interception:

1. First determine the most significant part of the area
2. Extract the feature map of the salient region
3. Identify the objects with deep networks

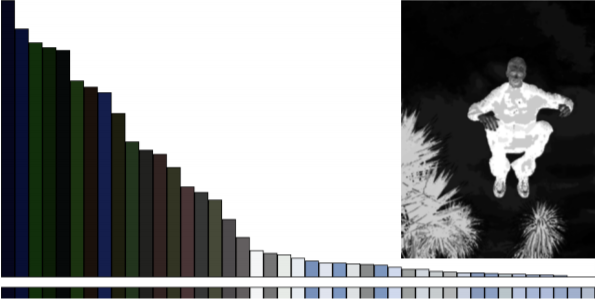
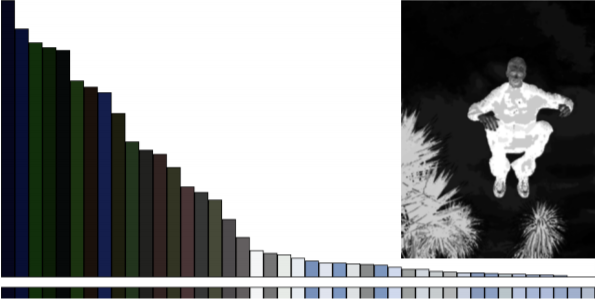
First of all, the significance test we use includes the following：

1. the contrast based on histogram statistics

The significance value of a pixel is determined by its contrast with the color of other pixels in the image (the distance in L\*a\*b space). Then the same color will have the same significance value, we can put together the same color values for calculation. However, both methods have a high time complexity, which can reach O(N^2) and O(N)+O(n^2). Therefore, we optimize to reduce the realism of the image within a reasonable range, in exchange for higher image processing efficiency.

1. The number of colors in each channel is quantized to 12 different values, and the number is reduced to 12^3.
2. Original image(left) and reducing the color(right) 
3. Discard the smaller frequency and choose the color that appears at high frequencies. And make sure the color covers more than 95% of the pixels, and the remaining 5% of the color is replaced by the nearest distance in the histogram.

Although we can efficiently calculate color histograms with color quantization, the quantization itself may be flawed (some similar colors may be quantized to different values) in order to reduce noise. We do a smoothing operation: the saliency value of each color is replaced by a weighted average of the saliency values of similar colors.



1. Image before smoothing(left) and image after smoothing(right)
2. the area-based contrast

We know that high contrast in adjacent areas is more likely to cause visual attention than high contrast in far areas. But calculating the pixel-level contrast introduces a spatial relationship that can be costly.

So, our idea is: we introduce an analysis method: regional contrast, combining spatial relationships with regional level contrast.

1. First divide the image into several regions
2. calculation area and color contrast
3. Use each region and other regions to weight the sum and define the significance value.

(The weight is determined by the spatial distance of the area, and the larger the distance, the smaller the weight)

Steps:

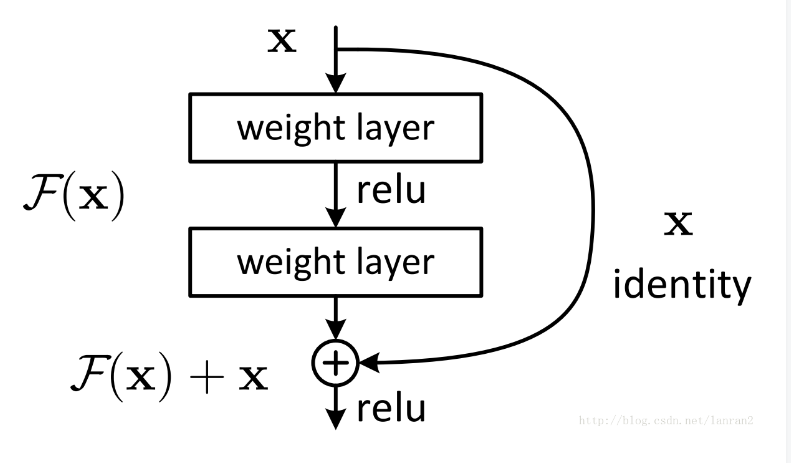
* 1. Split the image into several regions by selective search
  2. Create a color histogram for each region
  3. Calculate the significance value by measuring its color contrast with other areas of the image
  4. This can take into account spatial relationships, distance weight

1. selective search(left) and salient region(right)



Residual network part

The deeper the network extraction feature is, the more we use the network with deeper layers to train, the problem of gradient disappearance and network degradation will occur. We use the superposition of residual blocks to solve this problem.

1. It can ensure that the derivative of each layer of the neural network is greater than one when the x-bias is obtained, which avoids the hierarchical phenomenon of the product.
2. When the network handles the same mapping problem, just let F(x)=0, which simplifies the calculation problem that makes the output function equal to x.
3. Schematic diagram of Residual network

# Key results and achievements

## Key results and discussions

Over the last whole year, we have been devoting time on this project, which finally received a good result as a reward to our devastating hard word, and which, along with its innovation, is the most important reason why our revolutionized work was done by our five group members. We decided to work on this project because of the willingness and desire of changing the poor efficient way where commercials are publicized, and improving the experience of users where they are exposed to commercials. With innovation, devotion and responsibility of changing the world, we finally made it.

Our work is mainly in the laboratory room, and has written code for the saliency detection and deep residual network. After obtaining a large number of training sets and test sets on ImageNet, it was completed by a Titan X GPU for two days. The accuracy rate can reach more than 95%.

What we have achieved now includes:

* Accurately determination on significance of objects appearing in videos.
* Object recognition in a continuous video.
* Advertisement recommendation based on significant object appearing in the video.
* The first realize on combination of digital video processing techniques and commercial activities.
* Feature Detection
* Object Recognition
* Algorithm Improvement
* Software Realization
* System construction

These features are stable and robust for daily utilization both theoretically and technically, for the simple reason that our work has been testified with real video stream circumstances. These tests received a good result, which strengthened our confidence about the utilization of our work.

The sift algorithm and HOG processing we completed learning furthers our understanding more about computer vision, convolutional neural network and saliency detection. We analyzed the content of the video, using which to help corporations to advertise more precisely. Besides scientific value, we believe our work also has great commercial value to advertise better. During this period, the R-CNN processing method was learned, and the image recognition and classification can be achieved with high accuracy. After that, we understand and reproduce the method used to process the video called T-CNN algorithm. We have a certain understanding and basis for processing video.

Our achievement is revolutionary not just because of its comprehensive utilization of avant-garde techniques, but its further deep impact on the way where commercials are circulated and functioned. However, there is still going to be multiple issues to be achieved in the future to make our work being promoted and transmuted in the future so that our work can really improve users’ experience. What we have achieved now needs to be promoted both technically and practically. Next, we are looking for advertisement providers and video service provider, asking for further cooperation with market and our techniques.

Due to the limitation of the reimbursement policy, our team has not yet had a computer equipped with a high-performance GPU for the project. The model we built only runs on our personal computer. However, the performance of personal computers is low, which limited the size of our model.

We introduced a tremendously efficient regional contrast based salient object extraction algorithm, which simultaneously evaluates global contrast differences and spatial weighted coherence scores. These saliency maps are further used to initialize a novel iterative version of, for high quality salient object segmentation.

1. Salient region detection

We introduce a regional contrast based salient object extraction algorithm, which simultaneously evaluates global contrast differences and spatial weighted coherence scores. These saliency maps are further used to initialize a novel iterative version of GrabCut for high quality salient object segmentation.

We use the Histogram Contrast algorithm (HC), which based on the color statistical characteristics of the input image and detect it.

We used Region Contrast algorithm (RC) to combine spatial relations and regional level contrast together, the picture first divided into several regions, with each region and their regional contrast weighting and to calculate the regional significance

OBJECT RECOGNITION

First, make a detection of the image, first find out the ROI (Region Of Interest) in images, for each ROI area using “ROIAlign” make pixel correction, and then for each ROI use the design of Fast-CN classification framework to predict different instance belongs to right or wrong, and get the segmentation result of the image instance.

Later, use the mask encodes the spatial layout of the input target. And use the m\*m matrix to predict each ROI rather than using a vector, this ensures that the spatial information in the ROI will not be lost

## The achievement against activities and milestones

What we have done now is slightly little bit different from what we expected at the very beginning. We did not directly process videos. Instead, we preliminarily take out pictures from videos which highly reduces the difficulty of processing recognition. Secondly, we successfully made an application not only for demonstrating our effect; moreover, it can also can be an experimental achievement for daily use, which is a giant leap from theory towards reality.

Another point that is different from what we expected is that it did not met up with too much difficulties during developing however, finding ways for achieving this took a lot of energy. What we have achieved now includes:

1. Accurately determination on significance of objects appearing in videos.
2. Object recognition in a continuous video.
3. Advertisement recommendation based on significant object appearing in the video.
4. The first realize on combination of digital video processing techniques and commercial activities.

Our achievement is revolutionary not just because of its comprehensive utilization of avant-garde techniques, but its further deep impact on the way where commercials are circulated and functioned. However, there is still going to be multiple issues to be achieved in the future to make our work being promoted and transmuted in the future so that our work can really improve users’ experience. What we have achieved now needs to be promoted both technically and practically. Next, we are looking for advertisement providers and video service provider, asking for further cooperation with market and our techniques.

# Impacts

Scientific impacts: We make all use of the cutting-edge methods in deep learning through preprocessing, feature exaction and finally the advanced Residual network. It is a brand-new attempt on combining these methods together and evaluate its performance on the advertisement field.

Community impacts: As for community impacts, we put forward a subversive concept among advertising: contents related advertising strategy. We not only come up with it but also realize it on a software which could be used directly into the commercial area. In terms of presenting feedback from professors and students as well as feedbacks online, we believe that this idea will intrigue sea change in the advertisement area.

# Conclusions and recommendations

In the whole process of project, first we reviewed the previous paper on image segmentation, feature detection and object recognition specifically ICCV and CVPR critical breakthrough in these two years and some fundamental work before. Then we set up experiment environment(vs2017+opencv) and try on code implement. Then we focus on key frames detection and selective search which are two key problem we went through. After that, software development is smoothly processing.

For follow-on projects, we recommend improvements on Residual network for better accuracy and take how to extract key frame into consideration.

# Acknowledgments

Special thanks to our mentor Fu Lihua, who answered all our questions even though some of them may have been rather amateur. We couldn’t have completed our Keystone so well without her.

And let us not forget our dear dean professor Guo, who have just said goodbye to our FGX Honors College. This project wouldn’t even exit if not for his effort.

Last and not least, we must thank our teacher Ke Xiaoxing, who taught us basically everything about how to make our report & poster more professional.

Since our Keystone is software-based and the development tools we used were all open-source, we DID NOT NEED OR USE any budget, hence no budget provider.

# References

[1] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[2] Girshick, R. (2015, December). Fast R-CNN. In Computer Vision (ICCV), 2015 IEEE International Conference on (pp. 1440-1448). IEEE.

[3] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[4] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2980-2988). IEEE.

[5] Rosten, E., & Drummond, T. (2006, May). Machine learning for high-speed corner detection. In European conference on computer vision (pp. 430-443). Springer, Berlin, Heidelberg.

[6] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

[7] Shuai, H., Liu, Q., Zhang, K., Yang, J., & Deng, J. (2018). Cascaded Regional Spatio-Temporal Feature-Routing Networks for Video Object Detection. IEEE Access, 6, 3096-3106.

[8] Xu, W., Miao, Z., Zhang, X. P., & Tian, Y. (2017). A Hierarchical Spatio-Temporal Model for Human Activity Recognition. IEEE Transactions on Multimedia, 19(7), 1494-1509.

[9] Kalogeiton, V., Weinzaepfel, P., Ferrari, V., & Schmid, C. (2017, July). Action tubelet detector for spatio-temporal action localization. In ICCV-IEEE International Conference on Computer Vision.

[10] Mettes, P., & Snoek, C. G. (2017). Spatial-Aware Object Embeddings for Zero-Shot Localization and Classification of Actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4443-4452).

[11] Galteri, L., Seidenari, L., Bertini, M., & Del Bimbo, A. (2017). Spatio-temporal closed-loop object detection. IEEE Transactions on Image Processing, 26(3), 1253-1263.

[12] Zhu, H., Vial, R., & Lu, S. (2017, October). Tornado: A spatio-temporal convolutional regression network for video action proposal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5813-5821).

[13] Hou, R., Chen, C., & Shah, M. (2017, March). Tube convolutional neural network (T-CNN) for action detection in videos. In IEEE International Conference on Computer Vision.

[14] Kumar, R., & Batra, D. (2016, March). Pose tracking by efficiently exploiting global features. In Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on(pp. 1-9). IEEE.

[15] He, K., Zhang, X., Ren, S., & Sun, J. (2014, September). Spatial pyramid pooling in deep convolutional networks for visual recognition. In european conference on computer vision(pp. 346-361). Springer, Cham.

[16] Escorcia, V., & Niebles, J. C. (2013, December). Spatio-temporal human-object interactions for action recognition in videos. In Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on (pp. 508-514). IEEE.

[17] Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2015). Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3), 569-582.

[18] Rosten, E., Porter, R., & Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. IEEE transactions on pattern analysis and machine intelligence, 32(1), 105-119.

Budget:

Most of our budget was spent on acquiring hard drives in order to save all the training data we gathered from the Internet and generated by ourselves. The main expenses are listed as follows:

|  |  |  |
| --- | --- | --- |
| Items | Detailed Description | Total Price |
| external hard drive | for data storage | ￥3000 |
| reference books | for related data study | ￥1200 |
| USB flash drive | for data exchange purpose | ￥800 |

1. Budget