Deep Learning Algorithm for Picture Frame Detection on Social Media Videos

Fucheng Zheng
Dept of Electrical and Electronic Engineering
Auckland University of Technology
Auckland, New Zealand
fucheng.zheng@aut.ac.nz

Cheng Yang
Zyetric Technologies Ltd.
Auckland, New Zealand
robert.yang@zyetric.com
0000-0002-4654-5861

Peter Han Joo Chong
Dept of Electrical and Electronic Engineering
Auckland University of Technology
Auckland, New Zealand
peter.chong@aut.ac.nz

George Wang
University of Harvard
Cambridge, MA, USA
fwang@college.harvard.edu

G.G.Md.Nawaz Ali
Dept of Computer Science and Information Systems
Bradley University
Peoria, IL, USA
gga@g.clemson.edu

Patrick Lam
Zyetric Technologies Ltd.
Hong Kong, China
patrick.lam@zyetric.com

***Abstract*—This paper introduces a novel method for picture frame detection by using a deep learning algorithm. The detection aims to find the four vertices of multiple picture frames on social media videos. The detection model is based on Key-point RCNN (Region-Based Convolutional Neural Network). Although the Key-point RCNN is suitable for human key points detection, it does not perform well on the vertices detection of picture frames. In this research, a new picture frame (PF) branch is created to replace the Key-point branch of the Key-point RCNN. This PF branch includes more convolutional layers of the neural network and a feature pyramid network (FPN) structure which can extract more detail of features of picture frames. The experiment shows that the new PF branch significantly increase the accuracy. In 12 test videos, number of good performance videos are raised from 1 to 9 for the picture frame detection.**

Keywords—deep learning, key point detection, picture frame detection, computer vision

I. Introduction

Deep learning for computer vision is getting popular in the field of academic research. Significantly, the Convolutional Neural Network (CNN) [1] is a successful algorithm on object detection technique on images and video processing. From R-CNN [2] to Faster RCNN [3], SSD [4] and Yolo V4 [5], the object detection are more and more accurate and faster. In these years, CNN is also utilized for particular functions, such as Key-point RCNN [6] for human pose estimation, MODNet [7] for image matting. In this research, a novel CNN method is introduced to detect picture frames on social media videos.

This research focuses on a particular challenge for picture frame detection on social media videos. Unlike standard object detection, which outputs bounding boxes, this research finds the four vertices of all picture frames on social media videos. A video may include more than one picture frame. The picture frame angles are different because of the different angles of recording cameras. In addition, the picture frames may be occluded by a human being who is moving on a video. According to the author's experience, no research focuses on picture frame detection, even no picture frame dataset. However, because a picture frame shows a quadrilateral/polygon shape or a three-dimension angle of a rectangle on a video, there is some similar research for references.

Reference [8] reported to modify the inception V4 network to find four vertices of a polygon shape object on an image called PolyCNN. However, this PolyCNN could only detect one polygon object on one image. Reference [9] used polar coordinates to rotate the bounding boxes of object detection. The rotated bounding boxes are more suitable for the objects' shapes. However, the network was built for a two-dimension rotation. It does not build a 3D rotation system. Reference [10] modified the Faster R-CNN network to detect text with quadrilateral shape. The model has shown that it could handle small occlusion, but it does not test the significant occlusion. More detail of similar researches is presented in section II.

This paper introduces a novel method for picture frame detection, which detects four vertices of a picture frame. The method also detects multiple picture frames and handles the occlusion problem. The proposed method is inspired by Key-point RCNN, which finds human body key points. There are three main contributions to this research. Firstly, the Key-point RCNN is implemented for picture frame detection. Key-point RCNN was used for human pose estimation. It detects key-points of the human body and handles some key points occluded by other human bodies. In the picture frame detection, Key-point RCNN is used to find the four vertices' points, which can be the "key points" of a picture frame. This is a new application of Key-point RCNN. Secondly, the Key-point RCNN for picture frame detection is not very accurate because of the simple structure of the Key-point branch. The proposed method builds a feature pyramid network (FPN) structure onto the Key-point branch. This is called the picture frame branch (PF branch). The PF branch improves picture frame detection significantly. Thirdly, the Key-point RCNN produces false positives in picture frame detection sometimes. Most of the false positives have invalid picture frame shapes. A parallel criterion method is created to filter out these false positives. It uses the difference of angles between the opposite sides of a picture frame to decide whether a picture frame detection result should be deleted or not.

The rest of this paper is organized as follows. Section II provides an overview of the related work on picture frame detection. Section III proposes the picture frame detection theories. Section IV presents the experimental results and discussion. Also, this section presents a comparison between the classical Key-point RCNN and the Key-point RCNN with the PF branch. In addition, it shows how well the parallel criterion filter out the false positive detection. Finally, section

V concludes the paper with future work on picture frame detection.

II. Related Works

According to the authors' experience, the detection of picture frames in videos has not been addressed in the literature yet, while other object detection methods are potentially applicable. Picture frame detection can be inspired from key-points detection, polygon detection, quadrilateral regression and 3D rotation rectangle detection. These research fields have differences and similarities.

There are various methods detecting objects, many of which that involve the use of key-points detection. For example, [11] proposed a method to improve the human body parts and human pose estimation by designing multi-scale residual modules to learn multi-scale in-depth features and give more precise body key-points. Reference [12] proposed an key-points detection approach to efficiently detect the poses of multiple people in an image. It used a non-parametric representation to learn to associate body parts. Then, a global context, greedy parsing steps improved performance and maintain accuracy, while two branches of the sequential prediction process allow learning body part locations and their associations simultaneously. Reference [13] built a convolutional network that takes in semantic masks extracted by a mask-R-CNN to predict a set of key-points to analyse objects poses. Reference [6] proposed a Mask Point R-CNN to increase object edge detection accuracy, using key-point detection technology combination to construct the contour of a target edge. Reference [14] proposed a solution that reduced the number of incorrect object bounding boxes by using a one-stage key-point-based detector to detect each object as a triplet (rather than a pair) of key-points. Similar as other object detection, key-points can used to detect object poses. The difference is that it finds key-points (or poses) of objects, not considering the shape or size of the object.

The polygon detection can also be used to detect objects. Different from key-points detection, it considers number of vertices and boundaries of the objects. Reference [8] uses polygon prediction, proposing a deep learning approach to predicting the (vertices of the) polygons encompassing the objects. Reference [15] proposed a two-step process to detect objects. It generated instance masks using segmentation networks before a deforming network transforms polygons to fit better object boundaries. Reference [16] presented a research which used an image crop, iteratively produces vertices of the polygon outlining the object. In another research, [17] used a polygon-based classifier for fine-grained categorization using colour differences. However, polygon detection has disadvantages. the methods of [8] and [16] did not mention the performance under occlusion problem. Reference [15] model could fail when initialization was inadequate.

Rotational modelling is another approach that can be used for object detection. The difference is that it considers the shape and direction of objects. Reference [18] improved detecting boundary points through a spatial FFT-based filtering approach, which allowed for direction generation of low noise 3D surfaces. Instead of using Cartesian coordinates to identify point locations, [9] used polar coordinates to produce a simpler object representation model which could reduce the number of regression parameters required to find the object's shape. However, this research can only detect objects with rotation in two dimensions, while picture frame detection requires three-dimensional rotation data or polar coordinate data.

Another object detection approach is the quadrilateral regression algorithms. This approach also detects points which is like key-points detection. However, the difference is to use the rectangle properties to inference the vertices' location. Reference [19] used an end-to-end, two-stage quadrilateral regressing network archi-tecture Faster R-CNN to detect objects. A quadrilateral region proposal network was used to generates candidate panels, the research classifies the candidates and refine their shapes. Reference [10] proposed the similar network architecture for scene text detection. The difference was that the vertices' location was calculated from the bounding box. They train both a quadrilateral detection head and a rotated rectangle detection head. Reference [20] propose modelling text location through corner points detection, with the text body corners predicted using a Deep-CNN and then refined.

In the proposed method of this research, the aim concerning picture frame detection is to find the four vertices of the frame. In addition, one vertex can be occluded by human. The polygon detection and quadrilateral regression research did not mention the performance under occlusion problem. The polar coordinates or rotation modelling requires three dimensions coordination information. Therefore, using those methods may not perform well in picture frame detection. However, key-points detection can detect occluded points of human being. If the number of key-points are reduced to four, this method is adaptable to detecting picture frames. The picture frame detection is inspired by key-points detection.

III. Methodology

The inspiration of picture frame detection proposed in this paper comes from the algorithm of key-points detection, which can also be called Pose Estimation. Because it is used to detect key-points of human body including the head, hand, elbow, and other joints. The key-points in MS Coco data [21] is 17. The difference between key-points detection and the picture frame detection is that a picture frame has only 4 key-points which are the 4 vertices of a picture frame.

A. Key-point RCNN for picture frame detection

Key-point RCNN [6] is to find the object belonging to the specific target category and its exact position in the given image and to assign the corresponding category label for each key-point instance. The Key-point RCNN method based on Faster R-CNN is intuitive in concept, flexible and robust, and has fast training and computational output characteristics. Key-point RCNN tries to identify the areas of the image where the object may exist. Generally, image processing and feature extraction are the first tasks of the neural network, which is the foundation of the whole computer vision (CV) task. Therefore, this part of the network structure is called the backbone. In Key-point RCNN, the ResNet-50 FPN is using as its backbone network. After the backbone, the Region Proposal Networks (RPN) is to propose anchors with a high recall rate. All objects in the image belong to at least one anchor. These anchor boxes are a set of rectangular bounding boxes, which may contain objects. Then use these anchors for classification and localization refinement. After that, the RoI Align crops and aligns the human RoI for human key-point detection. The key-point branches produce the human key-

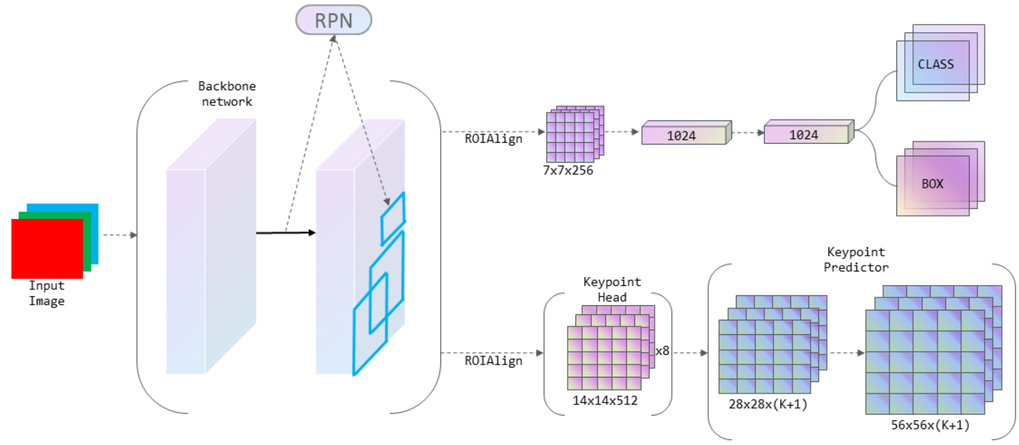


Fig. 1. The classical architecture of key-point RCNN.

point detection. It includes the Key-point Head and the Key-point Predictor. The structure of the Head includes eight convolutional layers of 14x14 of 512 channels. In the Predictor, the features are up-sampled two times, the resolution of 56×56 is finally output. Fig. 1 shows the classical architecture of key-point RCNN.

Different from human key-points detection, to implement Key-point RCNN to the picture frame detection, the outputs should be modified. The Key-point RCNN outputs the human body key-points (usually 17 points). However, picture frame detection requires the outputs of 4 vertices of a picture frame. Therefore, the outputs of the Key-point RCNN are modified to (4+1). In addition, from the experiments of different backbones, the Wide-ResNet-50 has the best performance for picture frame detection. In summary, there are two parts which different from Key-point RCNN for human pose estimation. They are the backbone of Wide-ResNet-50 and outputs of (4+1).

The modification of Key-point RCNN reaches the aim of picture frame detection. However, the experiment (in section IV) shows that the accuracy is not enough for industry application. To improve the accuracy, the Key-point Head and Predictor ("key-point" branch) should be modified for picture frame detection. This is discussed in the next section.

B. *The new Key-point branch (PF branch) for picture frame detection*

The proposed new key-point branch network is based on Mask Point RCNN[22], which improves the Mask-RCNN. It reported that in the "mask" branch of Mask-RCNN, the full convolutional network structure ignored the differences in spatial information between large-scale and small-scale reception domains. Therefore, they built a feature pyramid network (FPN) structure in the "mask" branch to improve the Mask-RCNN. The "key-point" branch in Key-point RCNN has the similar problem of "mask" branch in Mask-RCNN. The vertices of the picture frame is to locate at the edge of the target. Because the "key-point" branch lacks consideration of the reception domain of such a small scale as the edge of the target, the detection of the picture frame's vertices is easily be inaccurate. To solve this problem, a new Key-point branch for four vertices of picture frame detection is proposed which is called the picture frame branch (PF branch). The proposed PF branch is inspired by [22], which builds FPN into the "key-point" branch of Key-point RCNN. This is the difference between the Key-point RCNN and the picture frame detection neural network in this research.

In PF branch, firstly, the channel on each layer is increased to 512 from 256. Secondly, the convolutional layer is used for down-sampling of the feature maps rather than using the max-pooling layer. This because max-pooling only finds the main feature, but the convolutional layer has more parameters such as weights and bias to tune during the training. This can increase detection accuracy. Fig. 2 shows the new PF branch in Key-point RCNN. In Fig. 2, The backbone, RPN networks and the ROI Align are similar to Fig. 1, but the Key-point Head is different. After ROI Align, a map with the size of 28x28x512 is generated at beginning of the Key-point Head. Furthermore, the down-sampling of the feature maps is done through the convolution layer to 7x7x512. Then the deconvolution layer is used to enlarge the feature maps to 14x14x512 and finally to the original resolution of 28x28x512. In addition, two lateral convolutional layers following ReLU are used to transform features between the down-sampling and deconvolution layers, which builds up the FPN architecture. After that, in Key-point Predictor, a 1x1 filter is used to do the convolution on the output of the Head to change the feature map's channel to 28x28x(4+1). Finally, the feature maps are interpolated to 56x56x(4+1), which is the output of the picture frame vertices masks. The PF branch structure replaces the "key-point" branch of the Key-point RCNN. The whole model is the proposed picture frame detection method. This method is more suitable for multi-scale picture frame detection because of the FPN structure. Therefore, it increases the picture frame detection accuracy.

C. *Parallel criterion*

During the development of the above picture frame detection models to find four vertices points of a picture frame, the model was found that it detected some incorrect quadrilateral shapes along with the correct quadrilateral detection. The picture frame detection model detected something which was not a picture frame. These incorrect quadrilaterals are false positives. Because there are no conditions to limit the angle of four corners of a picture frame, the model cannot judge whether a result is correct or incorrect. To avoid this problem, a parallel criterion is proposed in this research. There are four sides in a picture frame. In a social media video, a picture frame captured by a camera that has a regular shape: the opposite sides (bottom-top sides and left-right sides) are almost parallel to each other. The difference of slopes (or slope angles) between the opposite sides is small. Therefore, these differences can be used to check whether these two sides are parallel or not. Suppose the slope of one side of a picture frame is (L). The slope is calculated below:

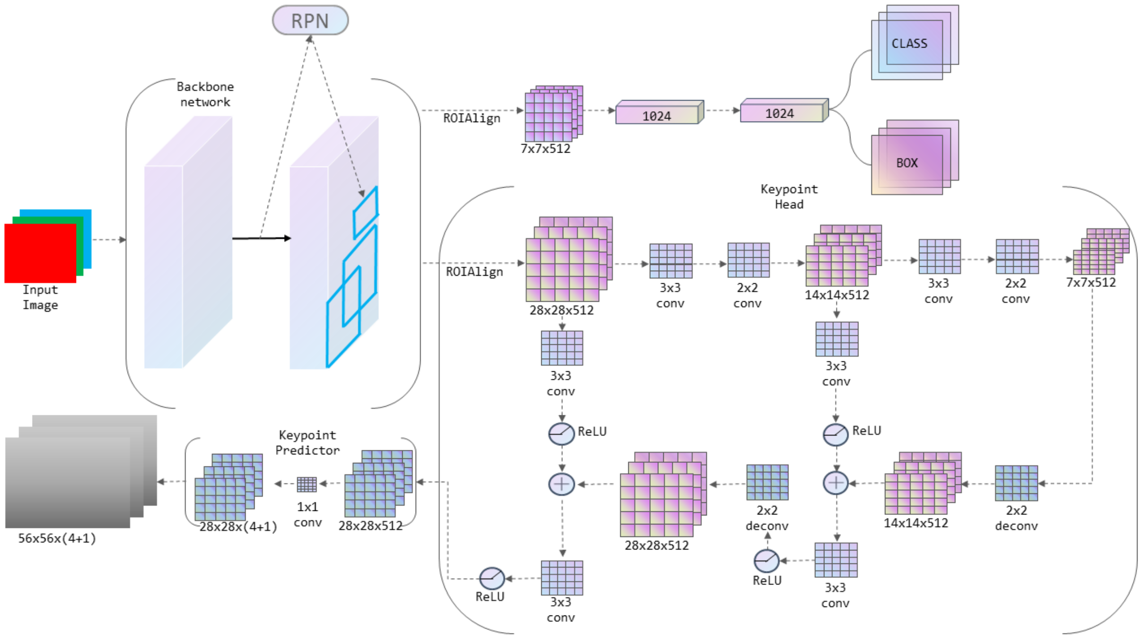


Fig. 2. The architecture of new Key-point branch (PF branch) for picture frame detection.

$$L = \left|\frac{y_1-y_2}{x_2-x_1}\right| \qquad (1)$$

where the (x_1, y_1) and (x_2, y_2) are the two vertices of a side of a picture frame. If x_2-x_1 =0, the slope is infinite, and the corresponding angle is 90 degrees.

```
Algorithm 1:

$A_{top}$ = top side's slope angle
$A_{bottom}$ = bottom side's slope angle
$A_{lift}$ = lift side's slope angle
$A_{right}$ = right side's slope angle

# Difference of the slopes angle
$D_{top-bottom}$ = abs ($A_{top}$ - $A_{bottom}$)
$D_{lift-right}$ = abs ($A_{lift}$ - $A_{right}$)

If ($D_{top-bottom}$ >9) or ($D_{lift-right}$ >9) then
    delete the picture frame detection
end If
```

The next step compares the slopes of the opposite sides of a picture frame detection result. The slope (L) is converted to its corresponding angle (A=arctan(L)). The comparison is the difference of the angles between the two opposite sides. If the angle is “zero”, the two opposite sides are parallel. Otherwise, they are not. If the angle is a small value, they are not perfectly parallel but can be accepted. However, if the angle is greater than a threshold, the detected quadrilateral is invalid. According to the experiment, the optimal threshold is 9 degrees (valid quadrilateral). If the angle is greater than 9 degrees, the picture frame detection is not accurate in the judgment of the human eyes (invalid quadrilateral), so the algorithm will delete the detection result. **Algorithm 1** shows the detail of the parallel criterion method.

In summary, the similarity between picture frame detection and Key-point RCNN is the architecture of backbone, RPN and ROI Align. The differences are the PF brand, output of 4 vertices and the parallel criterion.

IV. Experiment Results

This section introduces the data collection and the performance of the proposed picture frame detection theories. Picture frames images are collected. They have labelled four corners (vertices) of picture frames. This data set is for picture frame detection models training. The performance of the proposed methods is tested by twelve videos that have picture frames. The two methods (classical Key-point RCNN and PF branch model) are also compared. The function of the parallel criterion is also tested. Results are shown virtually because the picture frames on the videos are not labelled.

A. Data collection for training

The picture frame data is collected for the training of the Key-point RCNN model and the proposed (PF branch) picture frame detection model. Images include human beings and picture frames. It allows the picture frames to be occluded with the human body. Fig. 3 shows an example of an image and the label. The left picture frame is occluded by the human head, and it is the occluded picture frame. The right picture frame is the non-occluded picture frame. The red quadrilateral shows the label of the picture frames. 2600 images including picture frames are collected. 40% of images have non-occluded picture frames, and 60% images have occluded picture frames. 2400 images are collected for training, 200 for validation.

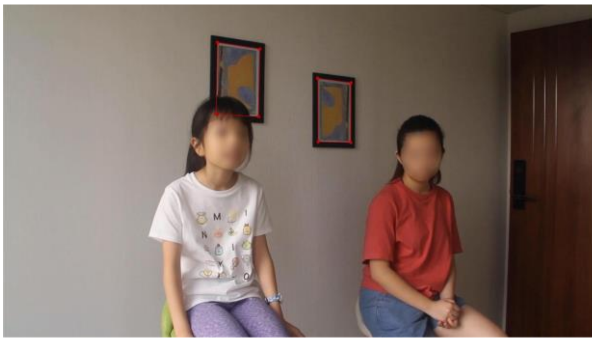


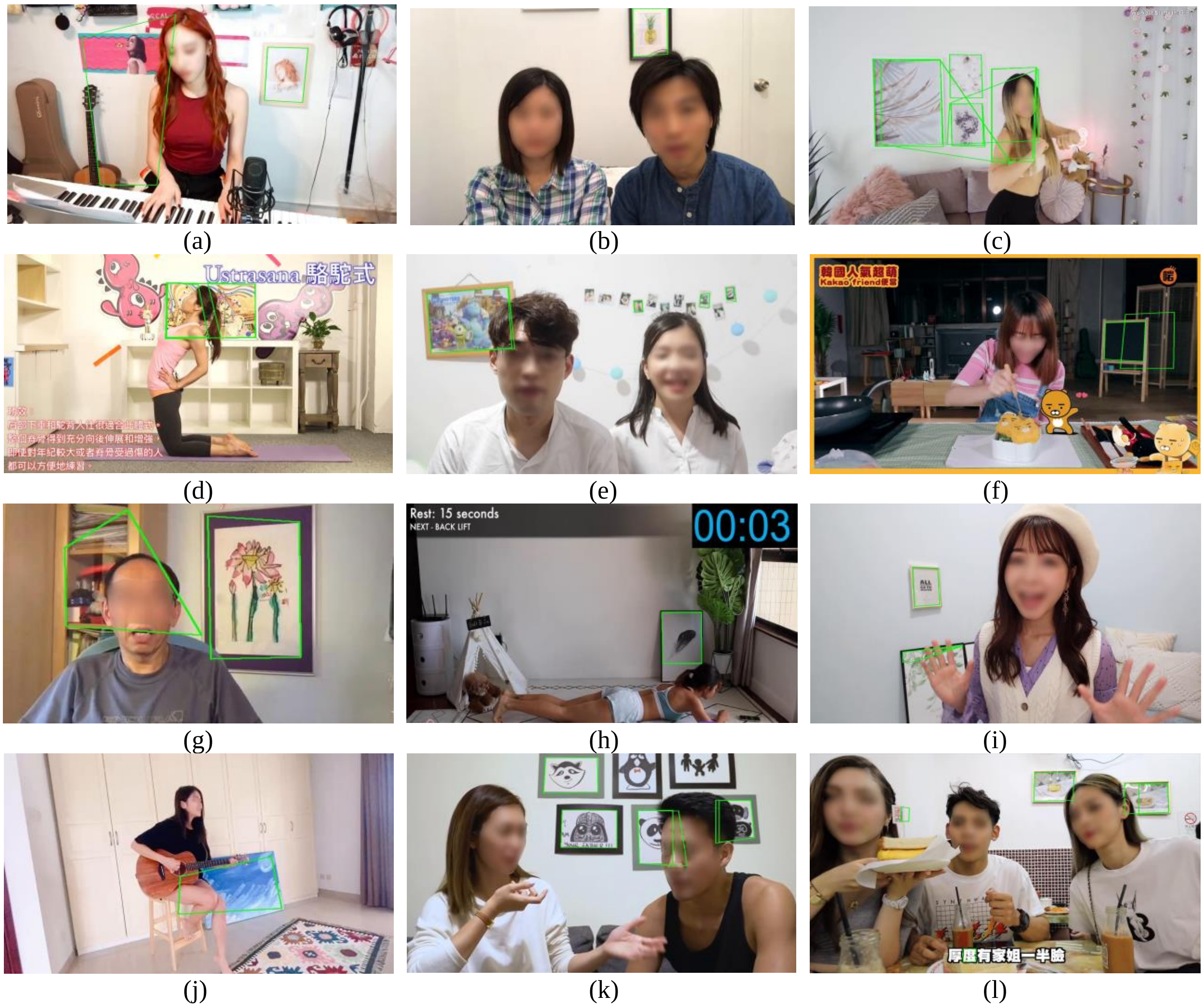
Fig. 3. An example of an image and the labeling.

B. Test and results

Twelve videos are collected to test the two picture frame detection methods. These videos are from KOLs, which cooperate with this paper authors. In this section, three models are tested and compared. The Key-point RCNN model for picture frame detection, the proposed picture frame detection model, and PF Branch for the picture frame detection + parallel criterion implementation. The virtual results are shown in Fig. 4, Fig. 5, and Fig. 6. The green quadrilaterals on these images are the detection results from the picture frame detection models.

In Fig. 4 (a), (c), (f) and (g), although some picture frames are detected well except video (g), the key-point RCNN model detects several noticeable false positive quadrilateral results because this model does not have a good classification ability for complex picture frames and chaotic backgrounds in these videos. Similarly, in Fig. 4 (d), (e), (k) and (l), due to chaotic backgrounds and multiples picture frames, inaccurate detection and repeating detection exist simultaneously. In addition, video (e) includes some small photos, and video (l) has a small flag. These small objects should not be detected in this research. Therefore, these detection results are false positives. In Fig. 4 (h), (i), and (j), the results of false positives are reduced. Because these videos have only simple pattern picture frames and simple background objects, however, some picture frames detections are inaccurate. In the video (i), Kol's head blocks one of the four key-points in the picture frame, which leads to inaccurate detection. Finally, only one video (b) has the successful detection of the four key-points, because the people, sofa and background in videos are straightforward. In addition, the human being in the video does not cover the picture frame too much. Therefore, this video has a good result.

Fig. 5 shows the results of Key-point RCNN with PF branch for the picture frame detection. According to detection results on videos (a), (c), (f), and (g), comparing to the classical key- point RCNN, the PF branch does not produce false positives and detects the picture frames more accurate. This means that it overcomes the complexity of the picture frames and the chaotic background. Even though video (c) appears an inaccurate picture frame detection, there are no false positives. In Fig. 5 (d), (e), (k), and (l), the problem of inaccurate detection and repeated detection is significantly decreased compared to the similar videos in Fig. 4. Video (k) has one more accurate result on the top picture frame but still has inaccurate detection on the two-button picture frames. Because these two picture frames are occluded two corners (vertices) by the KOL's head. In video (e), all small pictures are not detected, following the research requirement. However, the big picture frame on the left still has a repeated detection. In Fig. 5, videos (h), (i) and (j) have much better



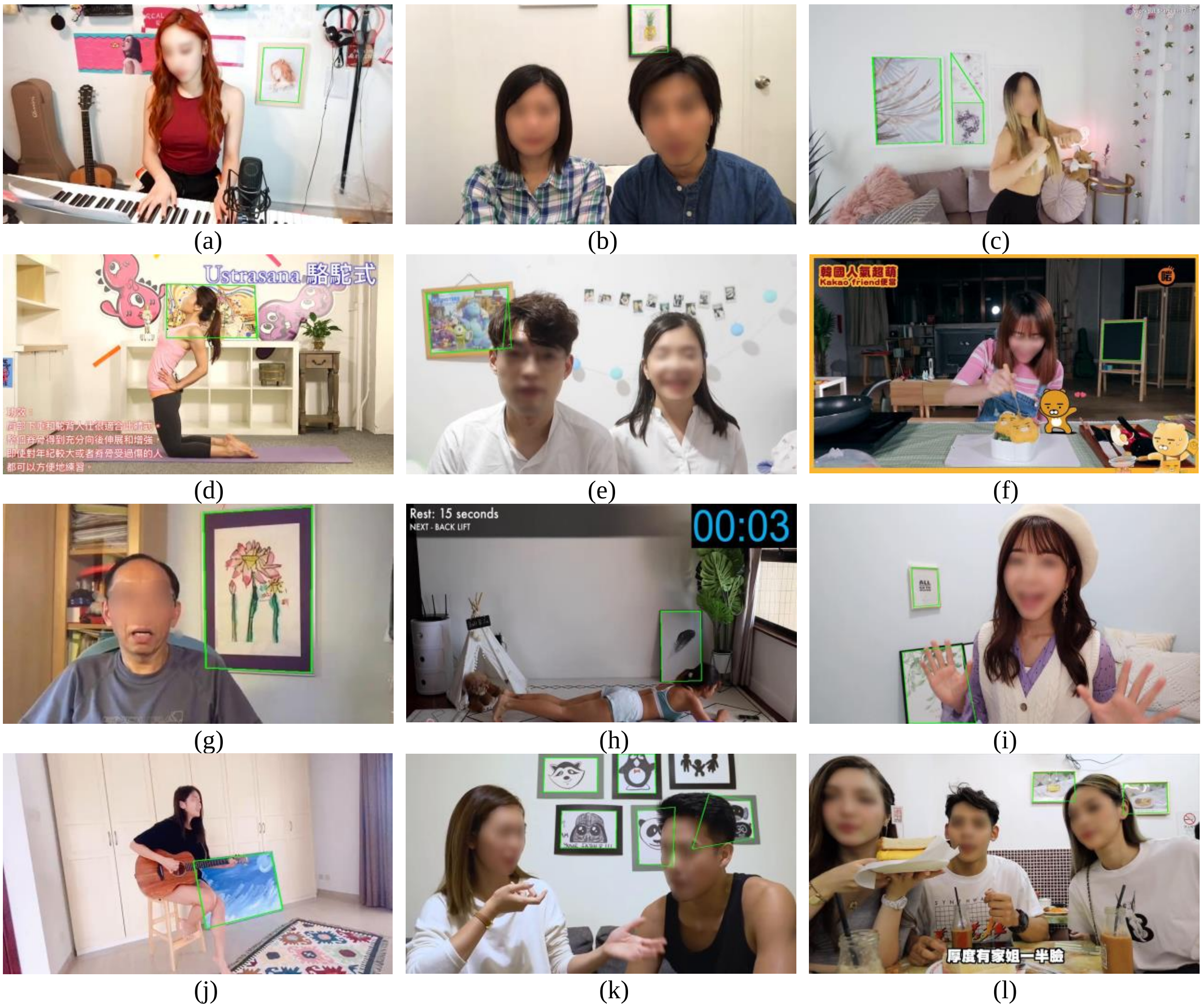
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 4. The classical Key-point RCNN results.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 5. The results of PF Branch for the picture frame detection.

results than the similar video in Fig. 4. In video (b), the picture frame is successfully and accurately detected. Finally, comparing to the Fig. 4, 8 videos (a, d, f, g, h, i, j and l) has better results in Fig. 5. This means the PF branch significantly increase the accuracy of the classical Key-point RCNN.

The performance of the PF Branch for the picture frame detection algorithm and the parallel criterion is shown in Fig. 6, which achieves very successful results on all videos. PF Branch has accurately detected the four key points of picture frames in most videos, so the parallel criterion improves a few videos with wrong results. They are videos (c), (e) and (k). The parallel criterion has successfully removed the repeated detection on video (e). On videos (c) and (k), because these inaccurate detection results do not show valid quadrilateral shapes, so they are deleted by the parallel criterion successfully. Furthermore, the parallel criterion has not removed the accurate picture frame detection result.

In summary, the parallel criterion successfully plays a secondary role in the algorithm. In most situations, it removes inaccurate detection and repeated detection. In addition, it does not remove the accurate detection. In the classical Key-point RCNN algorithm, inaccurate detection, false positive, and repeated detection may occur on the same video. In the test 12 videos, there is only 1 video having no problem. Conversely, there is no false positive result detected by the proposed picture frame detection method with PF Branch. The good performance video is increased to 9. PF Branch for the picture fame detection and parallel criterion performs the best results.

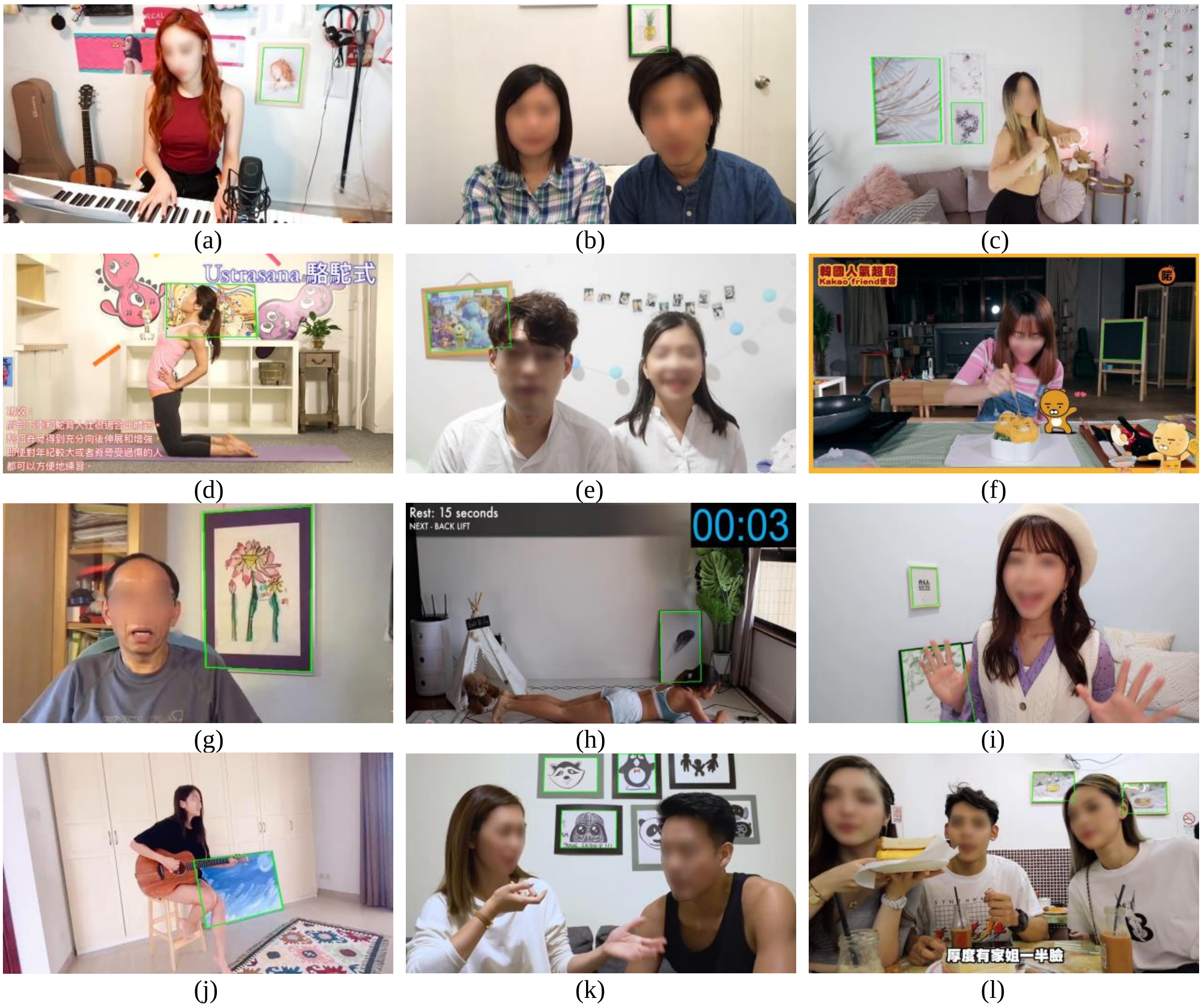
V. Conclusion

This research successfully implements the Key-point RCNN to detect vertices of picture frames on social media videos. The new PF branch is created and significantly increases the performance of the classical Key-point RCNN. The experiment shows that the new PF branch reduces the false positives of the picture frame detection. In addition, the picture frames are detected more accurate by the new PF branch than the classical Key-point branch. The parallel criterion with a 9-degree threshold can filter out most of the false positives on both classical Key-point RCNN and the Key-point RCNN + PF branch. In addition, some invalid shapes of picture frame detection results can also be removed.

The disadvantage of this method is the two vertices occlusion of a picture frame. According to Fig. 5 video (k), picture frames occluded by humans with two vertices are not detected well. The future work is to detect the picture frames which are occluded more than one vertex by other objects.

References

[1] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," International Conf. on Engineering and Technology (ICET), pp. 1-6, 21-23 Aug 2017. doi:



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 6. The results of PF branch for the picture frame detection + parallel criterion.

10.1109/ICEngTechnol.2017.8308186.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conf. on computer vision and pattern recognition, pp. 580-587, 2017.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, vol. 28, pp. 91-99, 2015.

[4] W. Liu et al., "Ssd: Single shot multibox detector," in European conf. on computer vision, Springer, pp. 21-37, 2016.

[5] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection,", 2020. arXiv preprint arXiv:2004.10934.

[6] W. Zhang, C. Fu, and M. Zhu, "Mask Point R-CNN,", 2020. arXiv preprint arXiv:2008.00460.

[7] Z. Ke et al., "Is a Green Screen Really Necessary for Real-Time Portrait Matting?,", 2020. arXiv preprint arXiv:2011.11961.

[8] N. Girard and Y. Tarabalka, "End-to-end learning of polygons for remote sensing image classification," in IGARSS IEEE International Geoscience and Remote Sensing Symposium, IEEE, pp. 2083-2086, 2018.

[9] L. Zhou, H. Wei, H. Li, W. Zhao, Y. Zhang, and Y. Zhang, "Arbitrary-oriented object detection in remote sensing images based on polar coordinates," IEEE Access, vol. 8, pp. 223373-223384, 2020.

[10] S. Wang, Y. Liu, Z. He, Y. Wang, and Z. Tang, "A quadrilateral scene text detector with two-stage network architecture," Pattern Recognition, vol. 102, p. 107230, 2020.

[11] R. Wang, Z. Cao, X. Wang, Z. Liu, and X. Zhu, "Human pose estimation with deeply learned multi-scale compositional models," IEEE Access, vol. 7, pp. 71158-71166, 2019.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE conf. on computer vision and pattern recognition, pp. 7291-7299, 2017.

[13] V. Pujolle and E. Hayashi, "Object 6 Degrees of Freedom Pose Estimation with Mask-R-CNN and Virtual Training," 2020.

[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE/CVF International Conf. on Computer Vision, pp. 6569-6578, 2019.

[15] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun, "Polytransform: Deep polygon transformer for instance segmentation," in Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 9131-9140, 2020.

[16] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-rnn," in Proceedings of the IEEE conf. on computer vision and pattern recognition, pp. 5230-5238, 2017.

[17] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "PBC: Polygon-based classifier for fine-grained categorization," IEEE Trans. on Multimedia, vol. 19, no. 4, pp. 673-684, 2016.

[18] C. Mineo, S. G. Pierce, and R. Summan, "Novel algorithms for 3D surface point cloud boundary detection and edge reconstruction," Journal of Computational Design and Engineering, vol. 6, no. 1, pp. 81-91, 2019.

[19] Z. He et al., "An end-to-end quadrilateral regression network for comic panel extraction," in Proceedings of the 26th ACM international conf. on Multimedia, pp. 887-895, 2018.

[20] K. Javed and F. Shafait, "Real-time document localization in natural images by recursive application of a cnn," IAPR International Conf. on Document Analysis and Recognition (ICDAR), vol. 1, pp. 105-110, 14th 2017.

[21] M. Kocabas, S. Karagoz, and E. Akbas, "Multiposenet: Fast multi-person pose estimation using pose residual network," in Proceedings of the European conf. on computer vision (ECCV), pp. 417-433, 2018.

[22] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," Sensors, vol. 20, no. 4, p. 1010, 2020.