Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 68-72, Ağustos 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 68-72, August 2020

Research Article

Analysis of an Image Recognition Method on Group Activities

Cemil Zalluhoğlu

[1] Hacettepe Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Ankara, Türkiye (ORCID: 0000-0001-8716-6297)

(Bu yayın 26-27 Haziran 2020 tarihinde HORA-2020 kongresinde sözlü olarak sunulmuştur.)

(**DOI:** 10.31590/ejosat.779063)

ATIF/REFERENCE: Zalluhoğlu, C. (2020). Analysis of an Image Recognition Method on Group Activities. *Avrupa Bilim ve Teknoloji Dergisi*, (Special Issue), 68-72.

Abstract

Recognizing group activity on still images is a very challenging problem. The difficulty in a distinction between foreground and background on images makes this problem more complicated than the problem of recognizing group activity on video due to the lack of spatial and temporal information. In this study, we examine the analysis of a still image recognition method on the Volleyball video dataset, which is collected for group activity recognition. We feed an additional mean image that is obtained from the previous and/or subsequent frames with the target image in order to analyze the temporal information gain. We aim to acquire temporal information from the mean images and to use it to train our method. As it is understood from the experimental results, our proposed method can get comparable results with the state-of-the-art video-based group activity recognition studies.

Keywords: Group activity recognition, Deep learning, Image classification

Grup Aktiviteleri Üzerinde Görüntü Tanıma Yöntemi Analizi

Öz

Sabit resimler üzerinde grup aktivitesi tanıma oldukça zorlayıcı bir problemdir. Resimler üzerinde ön ve arka plan ayrımını yapmak, uzamsal ve zamansal bilginin olmaması nedeniyle, bu problemi video üzerinde grup aktivite tanıma problemine göre daha zor kılmaktadır. Bu çalışmada grup aktivite tanıma için oluşturulmuş olan Volleyball video veri kümesi üzerinde, sabit resim tanıma tabanlı bir yöntemi incelemekteyiz. Kullanılan veri kümesindeki her bir video için bulunan hedef video karesine ek olarak, hedef karenin önceki ve sonraki karelerinden elde edilen ortalama görüntülerin resimler üzerinde zamansal bilgiye yaptığı katkının analizi incelenmektedir. Video grup aktivite tanıma problemlerinde sıkça kullanılan zamansal bilgi, ortalama resim kareleri üzerinden elde edilmekte ve resim tanıma yönteminin eğitim aşamasında kullanılmaktadır. Deney sonuçlarından anlaşıldığı üzere, önerilen yöntemimiz son teknoloji zamansal bilgiyi kullanan video tabanlı çalışmalar ile karşılaştırılabilecek sonuçlar alabilmektedir.

Anahtar Kelimeler: Grup aktivite tanıma, Derin öğrenme, Resim tanıma.

1. Introduction

In recent years, a dramatic increase has been observed in the field of computer vision research in studies on human movements and activities. The main reason for this is the increase in the use of the camera in daily life and accordingly, the image and video data obtained increases significantly. The need for automatic examination of this constantly increasing visual data has become important. The impact area of the studies carried out in this field includes a wide range of influences from security systems to entertainment industry.

Collective activity, also called group activities, can be defined as a sequence of movements performed jointly or interactively by more than two people. In order to model collective activities, it is necessary to recognize the individual movements of the individuals in the relevant group, as well as the spatial and temporal states of these movements relative to each other.

By its nature, group activity recognition has a more complex infrastructure than individual action recognition. In addition to individual actions, person-person interaction and group-person interactions are also very important in group activity recognition.

Besides, the problem of group activity recognition in still images is more difficult than that of group activity recognition in videos since it does not contain movement and temporal information. For this reason, the studies related to the problem of group activity recognition in still images in the field of computer vision are less than those of video-based studies.

This study is about the problem of group activity recognition in still images. Still image recognition method is applied and analyzed on an existing video group activity dataset [1]. In the proposed method, average images obtained using the previous and next frames of the labeled video target frame in the dataset are used to provide temporal information. The overall proposed method is shown in Figure 1. These images are used in the training phase together with the target frames. This proposed method achieves more successful results than most video-based studies in the literature.

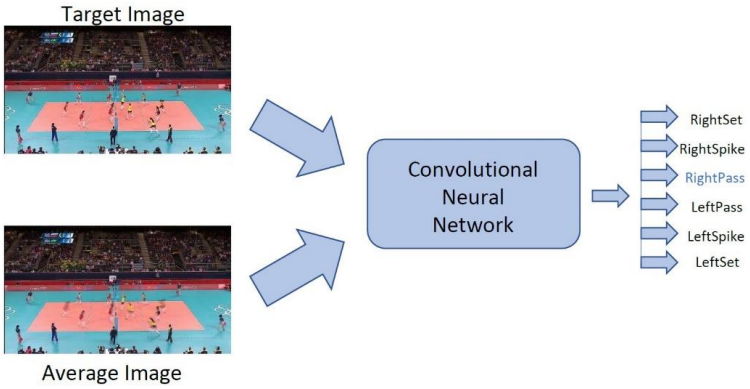

Figure 1. Analysis performed by Convolutional Neural Network (CNN) model-based image recognition method.

The group activity recognition problem is a new issue that has been studying in recent years. This problem is addressed through still images and videos. Group activity recognition on videos is more widely studied than group activity recognition in still images and studies on still images are less than the others. In this study, a comparative analysis is made between the proposed still image-based group activity recognition and video-based studies and the performance of this proposed method on the video datasets is calculated.

Ibrahim et al. [1] proposed multiple hierarchical long short-term memory (LSTM) methods to recognize group activities on videos. In this hierarchical architecture, the first LSTM architecture is used to capture individual actions, while the second LSTM architecture is used to achieve group activity dynamics in temporal domain. This model was further developed in [2] by adding an energy layer instead of the softMax layer that exists in the LSTM architecture. Qi et al. [3] proposes a recurrent neural network called stagNet that uses spatial mechanisms of attention and semantic graphics to recognize the collective activity. In order to model relationships in group activity recognition, [4] offers a two-level (person and stage) attention-based model. This model used two-stage Gated Recurrent Units (GRU) networks to provide temporal variability and consistency. From recent studies, Lu et al. [5] have designed a graphical convolutional neural network that investigates the interaction relationships in collective activities. In [6], they proposed a multi-stream spatial convolution neural network architecture that focuses on the regions of the person in both temporal and spatial channels to recognize collective activities on videos.

Studies related to group activity recognition problem on still images remain limited due to lack of dataset. In [7], which is one of the rare studies in this field, the only dataset in this field is proposed and interaction patterns are learned on this dataset to encode the relationships between people. Another study [8] has proposed a CNN-based two-stream architecture for combining RGB images and pose masks.

The image recognition problem is one of the most popular studies in the field of computer vision. Especially with the development of deep learning methods, the success of these studies increases even more. The methods mostly used in this field are AlexNet [9], VGG [10], Inception [11], and ResNet [12]. In this study, ResNet architecture is used for the image recognition problem.

The rest of the paper is organized as follows. Details of the method proposed are described in Section 2. In Section 3, the results of the experiment obtained by the proposed method are presented. Finally, we present the conclusions in the section 4.

2. Materials and Methods

2.1 Problem Definition

When a series of image is given, the group activity recognition method is defined as the estimate of the group activity classes for each image frame. The training set consists of $K_e = (R_i, S_i)_{i=1}^{N}$ with N samples. R_i is the *i*th training image frame and S_i is a finite number

from a set $S = \{1, 2, .. C\}$, of C group activity categories related to the category. In the testing phase, our aim is to match an unseen image frame with the correct one of the C group activities.

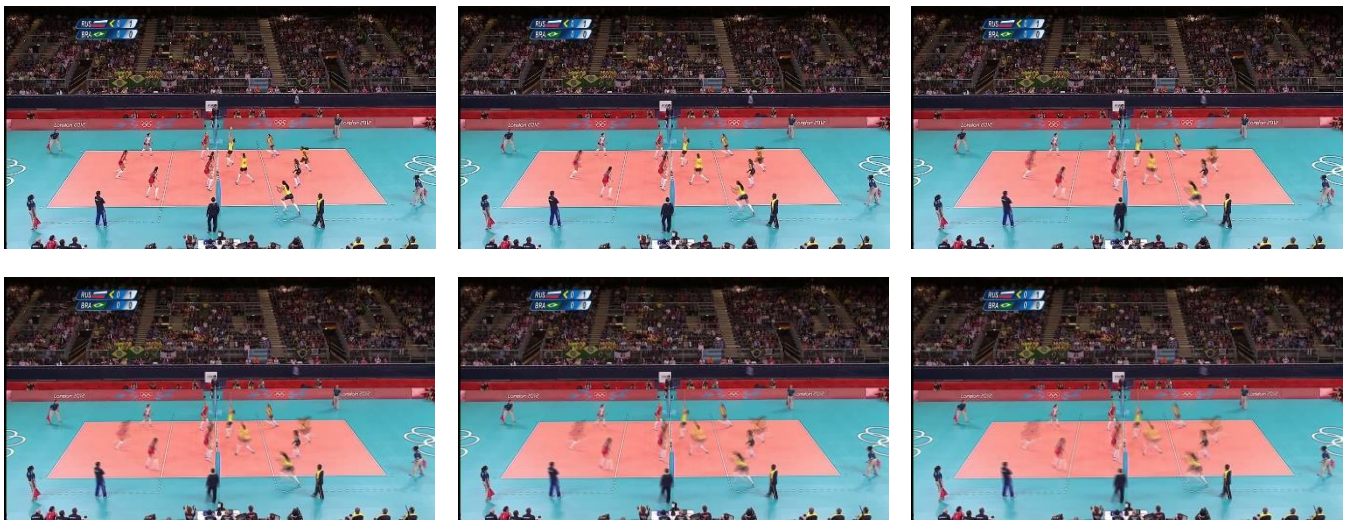

Figure 2. Average images created on frames before and / or after the target video frame. Sample images are left-to-right, *0_1_0, 1_1_0, 2_1_1* in the top row, and left-to-right *3_1_2, 4_1_3, 5_1_4* in the bottom row, respectively.

2.2 Recommended Model

As can be seen in Figure 1, we propose a CNN based image classification method. RGB image frames and averaged images are given as inputs to the method we recommend. The model first extracts the attribute information according to the given image frame. Then, class scores are calculated for the given image frame, $F_{c\in C}(R, W)$. In this formula, W represents the parameters that the method should learn. Softmax function is used to convert class scores into probability values on a class basis. The cross-entropy loss function (Eq 1.) is used to learn the parameters of the method.

$$\mathcal{L}_{act} = -\sum_{j}^{a} y_j \; log\left(\frac{e^{o_j}}{\sum_i e^{o_i}}\right) \tag{1}$$

In the equation above, a represents the number of activity classes, o_j is the output score of the jth collective activity class, and y_j represents the ground truth score of the given class.

As mentioned earlier, in addition to the target frame, average images are also used in the proposed model. Here, obtaining average images are defined as follows. "x_y_z" . x refers the number of images before the target frame is added to the average, y defines the target frame, while z refers the number of images after the target frame is averaged. 3_1_2 representation, for example, refers an image obtained as a result of the average of the target frame and three frames before the target frame and two frames after the target frame. Table 1. shows the experiments with the averaged images as well as the target frame.

3. Experiments

3.1 Dataset and Preprocess steps

In this study, Volleyball dataset [1], which is a video dataset, is used because of the lack of dataset related to group activity recognition problem in still images. This dataset contains a total of 15 videos, 1525 clips, and 41 image frames within each clip, and a target image frame labeled for each clip. This target image frame corresponds to the 21st frame in the video. In this dataset, there are six group activities in total: *spike*, *set*, pass, and the *left and right team* variants of these activities. In [1], the authors proposed a LSTM-based method that predicts the middle target (21st) image frame to determine the activity in each clip. In this approach, ten frames (five frames before and four frames after the target frame) were used. As in [1], 2/3 of the videos from each class is used for training and the remaining videos are used for testing.

In this study, this dataset is organized to define group activities in still images using target image frames instead of the videos. Inspired by the method proposed in [1], the following data augmentation method is used during the training phase in this dataset. As the preprocessing step, in addition to the target frames, new images are created by taking the pixel-based numerical averages from the

images taken from the previous and next frames of the target frame together with the target frame. Moreover, these created images are stored. Sample images are given in Figure 2.

3.2 Architecture and Application Details Used

In experiments, the 21st video frame is selected as the target video frame. Average image frames are used only in the training phase while only the target video frames in the test set are used in the test phase. For each image frame, attributes are extracted from the ResNet-18 [12] model and group activity classification is done through these attributes. Each method is run five times and the accuracy values are calculated by averaging these experiments.

For the proposed method, the learning rate is determined as 0.001 and the momentum as 0.9, and the learning rate is reduced by 0.1 per 10 epochs. This method is trained within 100 cycles. In the training process, the size of the training set is fed with four image frames in each period. As the data replication processes, the steps for cropping and rotating horizontally are randomly applied on all the images for the training set. In the testing process, only the central clipping is done. This study uses the previously trained ResNet-18 architecture on the ImageNet [9] dataset.

3.3 Experimental Results

In Table 1, the analysis of various combinations of the proposed approach is presented. It is observed that the results are improved when the average images obtained with the target frame are used. As a result of various experimental combinations, it can be observed that the success of the proposed methods decreases as the distance between the average images and the target frame increases. On the other hand, it is observed that the results obtained by using the average images together with the target frame during the training phase are better than the baseline study [1] in the video dataset. The best result is obtained by using the *1_1_0* method in which the target frame and the average of this frame and the previous frame are used together, and 68.48% accuracy is about 10% better than the video-based study [1].

Table 1. Results of experiments using images obtained with the average of a different number of frames

Method (x_y_z)	Accuracy
0_1_0	59.64
1_1_0	68.48
2_1_1	67.00
3_1_2	65.04
4_1_3	59.76
5_1_4	59.80

The proposed approach has been compared with state-of-the-art studies that define group activities using video processing techniques. The results are summarized in Table 2. In order to make a fair comparison, the same evaluation protocol is applied as [1]. The image classification approach, made using only target frames, provides about 8% better success than [1] using the complex video processing approach based on LSTM. The proposed method in which target frames and the average images are used together achieves a better result of about 2% than [13], an LSTM-based study that has achieved results on this dataset. However, it achieves comparable results with [6] that uses temporal and spatial information together with multi-flow convolutionary neural networks.

Table 2. Comparison of the proposed method with the results obtained in the Volleyball dataset.

Method	Accuracy
Two-Stage Hierarchical Model [1]	51.10
Image Classification (0_1_0)	59.64
SBGAR [13]	66.90
Proposed Method (1_1_0)	68.48
Region Based Multi Stream Model [6]	**72.40**

The confusion matrix of the proposed method is given in Figure 3. According to this matrix, most of the confusion that occurs between *set* and *pass* classes. There is a confusion in the *set* and *pass* activities that take place on both the left and right teams. The main reason for this situation is thought to be the scarcity of temporal information due to not using all video frames.

Figure 3. Confusion matrix obtained by the proposed method

4. Conclusion

In this study, an image-based classification method is analyzed on the video group activity dataset by selecting a target video frame by training these frames with before and after frames of the trained frames. With this analysis, it is observed that the proposed method achieves comparable results with video-based approaches with less effort and processing power. In this context, it is emphasized that in the video group activity recognition problem, an image classification method training can be done with the images obtained with the target frame and data augmentation approaches. As a result of the experiments carried out, it is observed that the success of this problem is increased by using the analyzed method in the video dataset. In addition to classification with a still image, it is observed that using average images containing partial temporal information improves the results even more. In future studies, it is aimed to use and to analyze the proposed method in video datasets in different computer vision fields.

References

1. Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2016). A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1971-1980).
2. Shu, T., Todorovic, S., & Zhu, S. C. (2017). CERN: confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5523-5531).
3. Qi, M., Qin, J., Li, A., Wang, Y., Luo, J., & Van Gool, L. (2018). stagnet: An attentive semantic RNN for group activity recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 101-117).
4. Lu, L., Di, H., Lu, Y., Zhang, L., & Wang, S. (2018). A two-level attention-based interaction model for multi-person activity recognition. Neurocomputing, 322, 195-205.
5. Lu, L., Lu, Y., Yu, R., Di, H., Zhang, L., & Wang, S. (2019). GAIM: Graph Attention Interaction Model for Collective Activity Recognition. IEEE Transactions on Multimedia, 22(2), 524-539.
6. Zalluhoglu, C., & Ikizler-Cinbis, N. (2019). Region based multi-stream convolutional neural networks for collective activity recognition. Journal of Visual Communication and Image Representation, 60, 170-179.
7. Choi, W., Chao, Y. W., Pantofaru, C., & Savarese, S. (2014, September). Discovering groups of people in images. In European conference on computer vision (pp. 417-433). Springer, Cham.
8. Akar, A., & Ikizler-Cinbis, N. (2019, September). Mask Guided Fusion for Group Activity Recognition in Images. In International Conference on Image Analysis and Processing (pp. 282-291). Springer, Cham.
9. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
10. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
13. Li, X., & Choo Chuah, M. (2017). SBGAR: semantics based group activity recognition. In Proceedings of the IEEE international conference on computer vision (pp. 2876-2885).