Mask Guided Fusion For Group Activity Recognition In Images[⋆]

Arif Akar[1] and Nazli Ikizler-Cinbis[2]

[1] Aselsan Inc., Ankara 06370, Turkey
arifakar@gmail.com
[2] Hacettepe University, Department of Computer Science, Ankara 06800, Turkey
nazli@cs.hacettepe.edu.tr
http://vision.cs.hacettepe.edu.tr/

Abstract. Recognizing group activities from still images is a challenging problem since images lack motion and temporal information that makes it easier to differentiate foreground from background. Nevertheless, images present rich spatial content that can be effectively leveraged for better feature representation and recognition. In this paper, we propose a two-stream convolutional neural network approach for group activity recognition. Our proposed approach is based on using person segment mask images to guide feature learning process. Our method is capable of inferring group relations without the need of bottom-up approaches and low-level annotations. To this end, we utilize three ways of fusing RGB and person segment mask feature maps. Experimental results demonstrate that person mask guidance provides a complementary learning process by outperforming previous methods with a large margin.

Keywords: Group Activity Recognition · Multi-stream Fusion · Person Segments.

1 Introduction

Group activity recognition is a challenging task in various ways. It shares similar challenges with human activity recognition problem such as occlusions, background clutter and change of appearance over time. However, recognizing group activities needs a more refined semantic understanding of the group scene, interrelations between members and their possible appearance features like view points. Although group activity recognition in videos is a considerably active topic of interest for computer vision community, group activity recognition in images is seldomly studied. It should be noted that recent successful methods [11], [10], [15] for group activity recognition in videos need complex hierarchical designs to explore semantic understanding of interactions with a bottom-up approach. Although bottom-up approaches are effective to make use of such contextual information; they require intense annotation work and complex design of

[⋆] This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) Research Program (1001), Project No:116E102

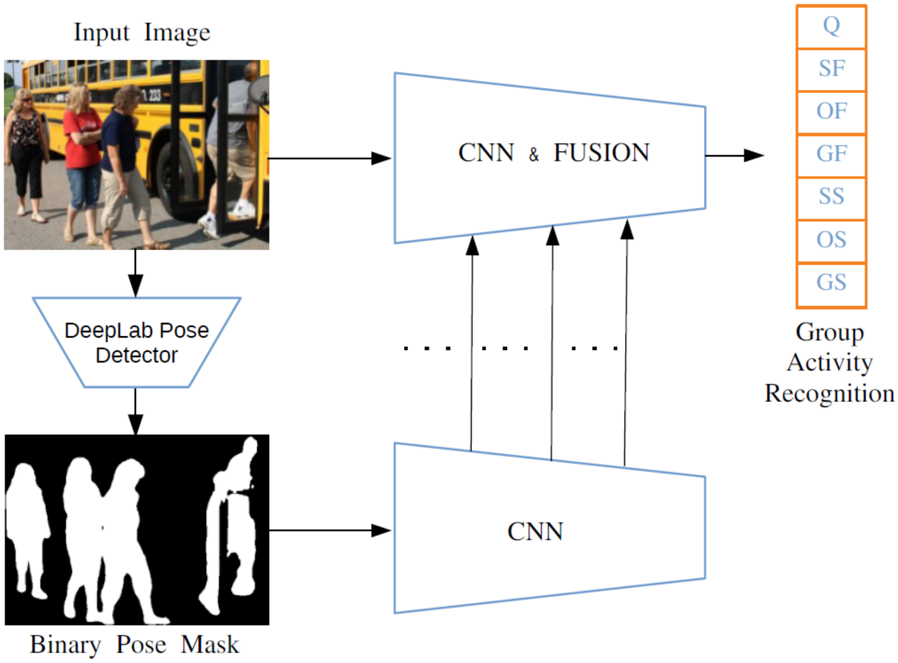

Fig. 1. The main framework of the proposed method.

low level components. This is partly because there is an exponential possibility of interactions between individuals and group activities show high variance in appearance.

In this paper, we tackle with the problem of recognition of group activities in images and we demonstrate an effective and simple method to deal with the aforementioned challenges. Our method improves spatial feature learning that possibly includes rich interaction and orientation features fiercely needed for group activity recognition. We intend to use the guidance of binary pose mask stream onto RGB stream to achieve a better representation of group activities in still images. Group activities can be inferred as a product of the interaction between individuals, their orientation and appearance information. Therefore, we explore ways of utilizing mask information to lead feature representation process by localizing some potentially significant parts of the image. Namely, our goal is to emphasize rich potentials of the image more than the remaining parts to capture valuable information for representing interactions and appearances within a group.

A simple illustration summarizing the proposed method is given in Figure 1. In our framework, for each input RGB image, a binary pose mask pair is generated using DeepLab[3] semantic image segmentation algorithm. Then, both RGB image and the binary pose mask are fed to separate CNN streams for feature extraction. Both RGB and Mask stream have identical CNN architectures. At intermediate steps, extracted feature maps from intermediate layers of Mask stream are fused into corresponding RGB stream to guide final feature map representation. We utilize three ways of fusing RGB and binary pose segment mask feature maps. Finally, fully connected layers at the end of the fused network is followed by the classification layer to compute group activity class scores.

The main contribution of this paper is three-fold. Firstly, we present a framework in which binary pose masks can guide feature learning from RGB images so that rich interaction between individuals and spatial appearance information can be extracted. Secondly, we evaluate different ways of fusing binary pose mask and RGB features in a convolutional network. Finally, we use SGD dataset [4] comprehensively to test our hypothesis against previous work and baselines.

2 Related Work

Group activity recognition research can be classified according to the type of source data. Research based on both video and image have been utilized in literature resulting in two distinctive approaches to this task. Although this paper focuses on activity recognition from still images, it is worth to mention video-based works in order to underline the similar motives behind both approaches. [8] is a GAN-based method in which generator can learn action codes from person level and group level features in a fusion scheme, while discriminator performs group activity recognition by validating the action codes as real or fake. In video-based research, RNNs are general preference to temporally reason over video frames. Authors in [2], propose a framework that is composed of fully-convolutional networks to extract a fixed-size representation and RNN to reason temporally for sequence of frames. In [11] and [10], LSTMs are utilized to capture temporal relations of the video and to represent and aggregate action dynamics. Another work using LSTMs is [15] in which authors designs a hierarchical LSTM network to model individual actions and interaction representations to reason on group activity. In [6], authors combine graphical models with RNN layers to leverage both rich spatial information and individual interactions. Inference algorithm reasons over individual estimates within a graphical model consisting of RNN nodes.

Group activity recognition using still images as the source of data has been limited by the availability of related datasets. One example is the work of Choi *et. al* [4], in which pose classifier, interaction classifier and group context classifier are learned using manually annotated RGB images. This bottom-up, hierarchical method makes a strong baseline for comparison. While this method uses rich low-level ground truth annotations including group, pose, orientation and interaction information during learning and detection (ground truth or poselet detector) information during test, we only use group information during the learning process.

Lack of crucial information in still images like temporal and motion components has been a natural force to find peculiar approaches for exploiting the most

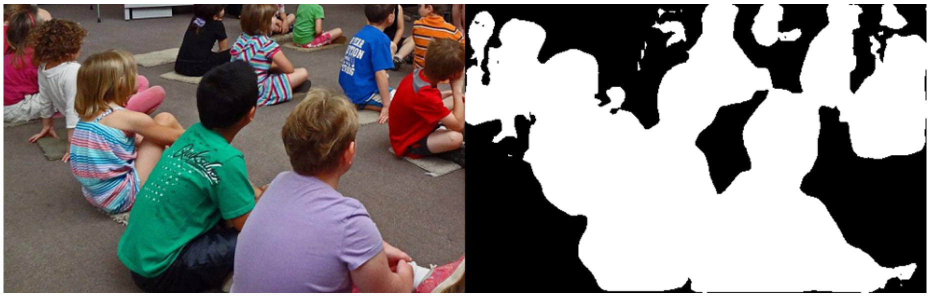

Fig. 2. A pair of RGB and mask inputs.

out of the spatial data. In [13], authors propose a method to compensate lack of temporal information. A Segnet based encoder-decoder framework is trained with segments of videos to learn temporal images hypothetically representing a sequence of frames. Then, these temporal representations are used to reason over still images for action recognition. Distinctively, [14] proposes a method to compensate for motion information missing in images. Unfortunately, both methods require manually annotated frames extracted from videos to recognize actions in images. Our method only relies on images during training without requiring extensive annotation burden.

3 Proposed Method

3.1 Problem Definition

Given a set of images I_N, the task of group activity recognition is defined as prediction of group activity classes for each image I_i. Let (I_i, Y_i) denotes the training set of RGB images and corresponding group activity class labels where Y_i takes labels from finite set, $L = \{1, 2, ...C\}$ for C classes. Then, we first obtain binary mask poses M_i for every image and form $(I_i^{H \times W \times 3}, M_i^{H \times W \times 1})$ input pairs. The final form of dataset can be denoted as $\{(I_i, M_i, Y_i) \mid I_i \in \mathbb{R}^{H \times W \times 3}, M_i \in \mathbb{R}^{H \times W \times 1}, Y_i \in \{1, 2, ...C\}$ for $i = 1, ...N\}$ where H and W denotes the resolution of the input. A pair of RGB and mask images is given in Figure 2.

3.2 Approach

We propose a CNN-based two-stream framework to fuse RGB and binary pose masks as shown in Figure 3. Streams of the network are fed with RGB images and its corresponding pose masks are extracted by a recently developed semantic image segmentation algorithm [3]. We use this segmentation tool for person class without any fine-tuning.

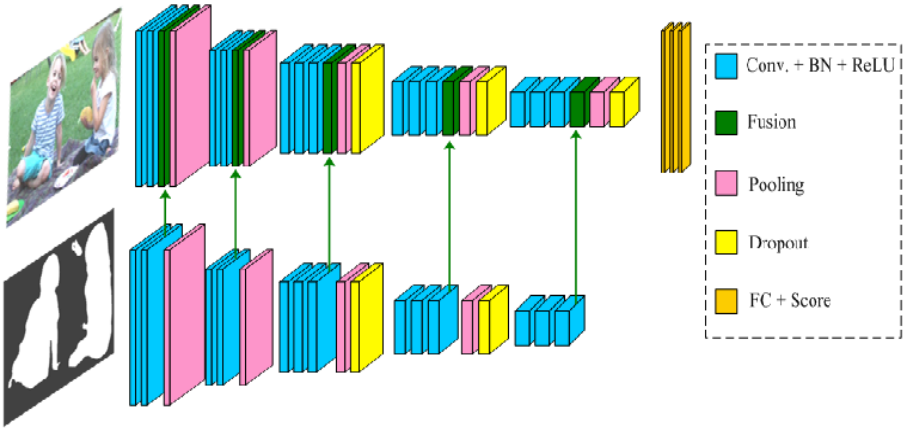

Fig. 3. Binary pose mask and RGB two-stream Convolutional Neural Network fusion architecture.

Our work is inspired by the encoder-decoder architecture proposed in [9] that uses depth images for semantic segmentation. Since we focus on classification task, our network streams consist of encoder architectures only. RGB and Mask streams include identical network architectures consisting of five sequential blocks of convolution, batch normalization and ReLU layers. Output of the encoder part is a combined feature map from two streams fused by one of the three fusion strategies. This is a high-level representation of a given image learnt from both RGB and binary pose masks. Finally, the combined feature map is fed to fully connected layers and a classification part to compute class scores.

The network extracts feature maps in the guidance of masks through fusion operation. Then, classification scores for input image I is computed as $f_c(I, W)$ for each class c where W refers to parameter set of the network and C is number of classes. In order to transform scores to a class probability distribution, *Softmax* function is used:

$$p(c \mid x, W) = \frac{\exp(f_c(I; W))}{\sum_{j=1}^{C} \exp(f_j(I; W))}, \tag{1}$$

To learn network parameters, we follow the optimization process used in [9] and [1], and use cross-entropy loss that is very useful for multi-class classification problems. Cross-entropy loss measures the Kullback-Leibler(KL) divergence between an input probability distribution and a target distribution. Since number of samples per each class show an unbalanced nature, median frequency balancing [7] is applied. This method balances classes of large sample numbers with smaller weights to compensate the unbalance of class size during training.

3.3 Network Architecture

We use a VGG-16 based network with 13 convolutional and 3 fully connected layers as shown in Figure 3. RGB and Mask streams are identical in terms of layer structure. We have 5 sequential blocks stacking convolutional, batch normalization and rectified linear unit layers (RELU). First two blocks have 2 x (64,128) weight layers respectively. Remaining 3 blocks have 3 x (256,512,512) weight layers. We keep the original pooling layers from VGG-16 network at the end of each block. At the end of the network, 3 fully connected layers reside as in the original VGG-16 model. We think VGG-16 is a sufficient and well-purposed architecture for exploring fusion methods; nevertheless, more complex architectures can also be explored.

3.4 Fusion Methods

We fuse features from both streams to explore the effect of guidance. Here, we follow the approach from [9], in which fusion of maps is performed by addition operation. There could be multiple ways to fuse feature maps from two-stream networks. In fact, fusion by element-wise addition operation is simply shown to have a stronger signal than single channel activations [9].

We utilize two of proposed fusion methods in [9], sparse fusion (SF) and dense fusion (DF) and we explore one additional fusion strategy called late fusion (LF). In sparse fusion, fusion is applied at the end of each (Conv.+Batch N.+ReLU) block. In Fig.4-a, sparse fusion is shown for second and third (Conv.+Batch N.+ReLU) blocks. Feature map after second Convolution-BatchNorm-ReLU layer from Mask stream, referred as MASK2-2, is added to the corresponding feature map from RGB stream and fed to pooling layer of second (Conv.+Batch N.+ReLU) block. Therefore, there are five fusion connections for five (Conv.+Batch N.+ReLU) blocks in SF experiment. Dense fusion is a more dense one in which features are fused after each layer in every block. As can be seen in Fig.4-b, number of dense fusion connections are two for second (Conv.+Batch N.+ReLU) block. In other words, there will be 13 fusion connections in total for DF experiment as number of layers in each block are (2,2,3,3,3) for all five blocks. In late fusion (LF), we only fuse features for once at the end of fifth block, before FC layers. Illustration of these fusion types are given in Fig.4.

4 Experiments

In this section, we present our results of the proposed approach for group activity recognition in still images.

4.1 Dataset

We use the Structured Group Dataset [4] that contains 600 images of groups of individuals generally encountered in daily life, such as at bus stop, cafeteria, classroom, conference, library and park. The dataset is rich in annotation-wise; group annotations, individual bounding boxes, 8 different viewpoints and

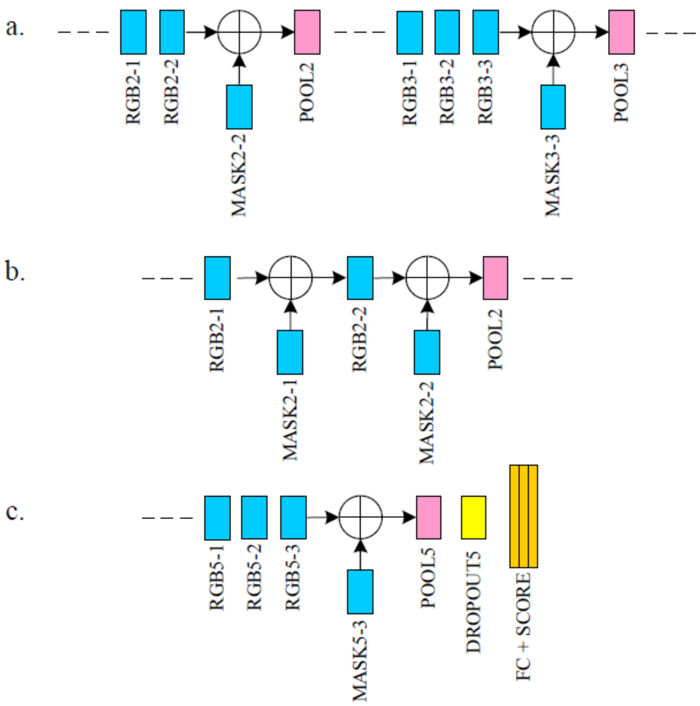

Fig. 4. Fusion methods for RGB and Mask streams of convolutional neural networks. a. Sparse Fusion at 2nd and 3rd (Conv.+Batch N.+ReLU) block b. Dense Fusion at 2nd (Conv.+Batch N.+ReLU) block c. Late Fusion at 5th (Conv.+Batch N.+ReLU) block are shown.

individual poses (standing, sitting on an object, sitting on the floor) are provided for each image. There are 7 labeled group activities: queuing (Q), standing facing-each-other (SF), sitting on an object facing-each-other (OF), sitting on the ground facing-each-other (GF), standing side-by-side (SS), sitting on object side-by-side (OS) and sitting on the ground side-by-side (GS). Since there can be multiple groups in a single image, 600 images contain 1743 group bounding boxes in total. We exclude 19 erroneous ones that have no person or only one person with no collective activity. This results in 1724 group images. It is also stated in [4] that similar amount of group images are removed as being outliers but no information on these images was disclosed in detail.

We split 1724 images into 80%-20% train-test split. Training set is augmented by flipping all 1379 images horizontally, resulting in 2758 images in total. Test split contains 345 images. The original work [4] also performed augmentation by flipping operation on the whole dataset, of which we only did for training part.

4.2 Training

We train all models with a batch size of 8 for 250 epochs (86250 iterations in total) with random shuffling at each epoch. All models are trained end-to-end with stochastic gradient descent (SGD). Learning rates were initially assigned between [0.0003, 0.0005] and decayed 10% at every 20 epochs. All networks are initialized with the standard VGG-16 model [16] pretrained using ImageNet [12].

Pose masks were generated using a DeepLabV3 [3] pretrained on Inception [5] using MS-COCO and VOC2012. Next, we transform all mask images into binary masks by filtering out every detected segments except for human segments. Mask images were resized to the resolution of 240x320 and paired with RGB images.

4.3 Baselines

The closest previous work that we can compare our method is [4], where authors apply a bottom-up solution to recognize group activities by learning individual poses, interactions and group context classifiers. Throughout their learning process, they make use of ground truth information including individual poses, individual bounding boxes, view points and group annotations. Unlike their method, our method uses only group annotations.

Our interest is to find out whether pose masks provide strong signals in emphasizing group interaction inference by masking out irrelevant surrounding context. Therefore, we also train baselines using only RGB-stream or mask-stream to evaluate the effect of the fusion.

Table 1. Accuracy comparison with previous work for group activity recognition on Structured Group Dataset [4].

Method	**Accuracy**
Choi[4] (Poselet Det.)	52.7
Choi[4] (Ground Truth Det.)	64.9
VGG-16 (RGB-only)	60.3
VGG-16 (Mask-only)	64.6
Dense Fusion	64.6
Late Fusion	65.8
Sparse Fusion	**70.4**

4.4 Results

We report experimental results for group activity recognition in Table 1 and in Table 2. Table 1 shows the overall accuracies of the baselines and related

Table 2. Class-wise average precision, recall and F1 scores over the Structured Group Dataset.

Method	P/R	Q	SF	OF	GF	SS	OS	GS	Avg
Choi[4] (Ground Truth Det.)	Prec	41.86	55.78	62.48	60.19	39.08	53.85	37.65	50.13
	Recall	27.48	64.55	65.56	65.00	21.33	40.86	26.52	44.47
	F1	33.18	59.85	63.98	62.50	27.60	46.46	31.12	47.13
Our method (SF)	Prec	58.33	68.92	72.91	83.33	45.28	73.13	89.47	**70.19**
	Recall	38.89	66.23	87.50	71.42	58.53	72.06	58.62	**64.75**
	F1	46.67	67.55	79.54	76.92	51.06	72.59	70.83	**66.45**

work, whereas Table 2 shows the classwise average precision, recall and F1 score comparisons between [4]'s best model and our best model.

As can be seen in Table 1, RGB-only model has the lowest accuracy amongst all of our models, whereas it still produces more accurate results than [4]'s model that operates on poselet detections. When using ground truth person detections, [4] method outperforms RGB-only CNN. Mask-only model is nearly on par with [4]-GT model; and this shows that person mask images can be a rich source for group activity recognition. It can be stated that dense fusion do not add much to learning more representative feature maps; probably due to the saturation in addition of similar inputs. This result also confirms findings in [9] where similar level of saturation observed for dense connections.

Late fusion of the RGB and pose mask streams seems to slightly increase the recognition performance; indicating that these two streams indeed carry complementary information, and pose masks extracted this way can be used as a guidance for focusing on the foreground and inferring collective group activity.

It is remarkable that sparse fusion of mask stream is able to improve group activity recognition accuracy with high margin. This result is interesting in two ways. Firstly, mask-guided fusion can lead network to learn high-level representative features within an end-to-end framework. This can help improvement of learning process for similar vision tasks even when a large-scale annotated data is not available. Secondly, the result implies that an optimum fusion method as in the case of sparse fusion is implicitly possible and can be searched in a finite architecture search space.

5 Conclusion

We have proposed a CNN-based two-stream fusion network to develop rich high-level representations of group activities. Our method does not require any low level annotations except for group activity label. In contrast to related work on the subject, our method directly infers interactions using binary person pose mask guidance. Experimental results show that mask guidance is complementary to learning feature maps from RGB stream, yielding superior recognition performance over bottom-up, annotation-heavy approaches for group activity recognition in images. Mask-based fusion can be applied to other tasks that need interaction inference for a better scene understanding.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561 (2015)
2. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Conference on Computer Vision and Pattern Recognition (2017)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
4. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Discovering groups of people in images. In: ECCV (2014)
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint pp. 1610–02357 (2017)
6. Deng, Z., Vahdat, A., Hu, H., Mori, G.: Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: CVPR (2016)
7. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2650–2658 (2015)
8. Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Multi-level sequence gan for group activity recognition. arXiv preprint arXiv:1812.07124 (2018)
9. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: ACCV (2016)
10. Ibrahim, M.S., Mori, G.: Hierarchical relational networks for group activity recognition and retrieval. In: ECCV (2018)
11. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: CVPR (2016)
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
13. Safaei, M., Balouchian, P., Foroosh, H.: Ticnn: A hierarchical deep learning framework for still image action recognition using temporal image prediction. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 3463–3467. IEEE (2018)
14. Safaei, M., Foroosh, H.: A zero-shot architecture for action recognition in still images. In: 2018 25th IEEE International Conference on Image Processing (ICIP) (2018)
15. Shu, T., Todorovic, S., Zhu, S.C.: Cern: confidence-energy recurrent network for group activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)