Multi-modal Pyramid Feature Combination for Human Action Recognition

Carlos Roig, Manuel Sarmiento, David Varas, Issey Masuda,
Juan Carlos Riveiro and Elisenda Bou-Balust
Vilynx
(carlos,manuel,david.varas,issey,jc,eli)@vilynx.com

Abstract

Accurate human action recognition remains a challenging task in the field of computer vision. While many approaches focus on narrow image features, this work proposes a novel multi-modal method that combines task specific features (action recognition, scene understanding, object detection and acoustic event detection) for human action recognition.

This work encompasses two contributions: 1) The introduction of a feature fusion block that uses a gating mechanism to perform attention over features from other domains and 2) A pyramidal feature combination approach that hierarchically combines pairs of features from different tasks using the previous fusion block. The richer features generated by the pyramid are used for human action recognition. This approach is validated using a subset of the Moments In Time dataset, resulting in an accuracy of 35.43%.

1. Introduction

Action recognition -and in particular, human action recognition- is foreseen as a key cornerstone in domains such as surveillance or video understanding, which has raised its market demand. However, despite recent advances in the field [1], accurate human action recognition still remains one of the most challenging problems in computer vision.

One important aspect of current human action recognition research is that most studies concentrate on human action feature representations but lack the capacity to detect and understand the contextual information associated with these actions. To alleviate this issue, a multi-modal system that combines action, scene, object and audio features is introduced in this work.

This paper addresses the multi-modal human action recognition task in videos by detecting people and using audio and visual information to leverage action features. To achieve this, a solution that combines the information of four separate architectures (action, scene, objects and audio) is presented in this paper. This combination is performed using a Pyramid Feature Combination architecture as it can be seen in Figure 1.

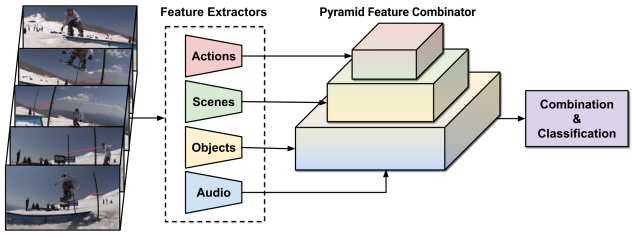

Figure 1. System overview. The system presented in this paper starts with a series of feature extractors that generate task-specific features. Features are combined using a pyramid structure, which combines features in a hierarchical manner. The features enhanced by the pyramid are finally used for human action recognition.

The main contributions of this paper are:

- A fusion block that uses a gating mechanism to perform attention over domain-specific features in order to transform them for further combination with features from a different domain.
- A novel pyramid approach that hierarchically combines pairs of features using the previous fusion block. The bottom-up pathway combines the features resulting in a rich action feature while the top-down pathway uses these richer features to leverage the task-specific features at every level of the pyramid. Finally, a lateral combination of the features is used to fuse each level feature and perform the human action classification.

This paper is organized as follows: Section 2 explains the feature extraction method and the proposed block for combining pairs of multi-modal features. Then, in Section 3, the pyramid feature combination approach is presented, detailing how it works and how it can be implemented. Section 4 describes the experiments performed on the Moments In Time dataset. Finally, Section 5 draws key conclusions based on different solutions proposed in this work.

2. Multi-modal Action Recognition

In this section, a system that effectively combines multi-modal video features for human action recognition is presented. First, a method that uses action features for human action classification is described in Section 2.1. Then, in Section 2.2, a temporal attention system that efficiently reduces feature dimensionality along the temporal axis is described. Finally, in Section 2.3, a block for combining pairs of features from different domains is proposed.

2.1. Action Recognition Feature Extraction

The action recognition system estimates the most relevant action observed along each video. As the scope of this work is the analysis of human actions for real case applications, a method that extracts and analyzes features associated with human detections is presented in this section.

The proposed multi-modal video action recognition method is composed by three main blocks. First, a detection module extracts human detections from each video to detect whether a human action may take place or not. Then, a set of convolutional neural networks are used to extract visual and audio features from the video and a temporal attention is applied to efficiently reduce the dimensionality of these frame level features. Finally, features from different domains are combined using a Pyramid Feature Combination Network to classify the human action that takes place at each video.

As a first step towards human action recognition, the human detector extracts bounding boxes associated with people present in the video. A network trained on the COCO dataset [10] is used for this purpose. Videos that do not contain any human detection are discarded because the scope of this work is detecting actions performed by humans.

As backbone action feature extractor for the analysis of videos with human detections, the *Inflated 3D ConvNet* (I3D) [2] trained with Kinetics [8] has been adopted. This has proven to be one of the most powerful architectures for trimmed action recognition and has also shown good performance in untrimmed action recognition scenarios [5].

Actions are predicted with short video clips formed by 64 frames. Each clip is passed through the I3D Network [2] up to an intermediate layer (*mixed_5c*) to extract image level feature maps of dimensions 8x7x7x1024. Note that at this layer, a temporal pooling has already been applied to the initial set of images. Finally, a GAP is performed in order to obtain a single vector representation for each temporal unit, resulting in a 8x1024 feature map. This feature map gathers the information about actions in the video. Although this feature vector may directly be used for action classification, the purpose of this work is to efficiently enrich these features using other visual and audio features.

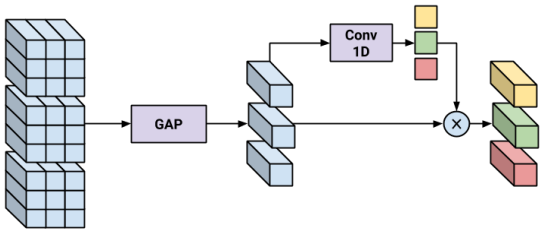

Figure 2. Temporal attention. The temporal attention block performs a pooling operation to reduce the dimensions over the spatial axes of the feature map, followed by a Conv1d to generate an attention score for each feature vector.

2.2. Temporal Attention

As the amount of relevant information carried by each frame of the video may vary, a temporal attention model is used to efficiently capture this information of the input data.

Specifically, the feature representation of the d_{th} domain (x^d) is obtained using a soft-attention based temporal pooling, which is formulated as:

$$x^d = \sum_{k=1}^{N} \alpha_k^d x_k^d \tag{1}$$

where N is the selected number of images per segment, α_j^d is the attention weight associated with frame j of the video and $\sum_{j=1}^{N} ||\alpha_j^d||^2 = 1$. The weights of the attention are computed as:

$$\alpha_k^d = \frac{(w_A^d)^T x_k^d}{\sqrt{\sum_{j=1}^{N} (\alpha_j^d)^2}} \tag{2}$$

where w_A^d is the 1x1xC attention kernel of the d_{th} domain.

The proposed temporal attention module is presented in Figure 2. First, a Global Average Pooling (GAP) is performed over the N feature maps in order to represent each image using a single feature vector of dimensions 1x1xC. Then, the attention weights are computed as presented in Equation 2 performing a 1D convolution over the channel dimension. Finally, these scalars are used to weight their corresponding feature vector and generate a 1x1xC representation of the video.

2.3. Multi-modal Feature Combination

Due to the nature of actions, which are usually related to specific scenes and sounds (i.e. fishing, snowboarding, cooking), a feature fusion method that takes into account scene and audio features to enhance action recognition is proposed in this section.

As it has been shown in previous works [7], using features from different nature may boost the performance of a single classification task. Some examples of feature combination are shown in [13], where the information of objects

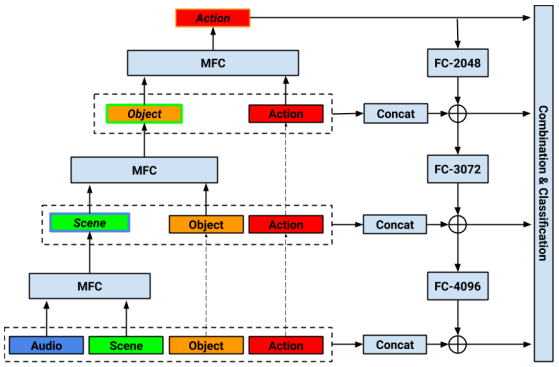

Figure 3. Pyramidal feature combination. The bottom-up pathway (left) uses the Multi-modal Feature Combiner (MFC) block to combine pairs of features. The top-down pathway (right) combines the features at each level to use them for classification.

that may be related with the scene is used to leverage scene recognition.

In this work, scene and audio features are used to complement local features extracted from human actions along videos. Intuitively, the presence or absence of certain multi-modal properties of the video may be used to infer a possible subset of actions that can take place. For example, a kitchen appearing in the video or the sound of frying something in a pan indicate that cooking is a possible action whereas the probability of snowboarding is reduced.

The proposed multi-modal fusion block is based on both a combination and a trainable gating of the features. In this work it is assumed without loss of generality that the pair of features x and y fused by this block belong to two different domains d_i and d_j respectively. First, a fully connected layer transforms features from d_j into pseudo features of d_i and projects them to the same dimensionality. Then, as both feature vectors belong to the same domain, they can be combined. A soft attention mechanism is used to weight the contribution of each feature before this combination. This mechanism helps to select which feature is more relevant at each part of the video. The weighted features are combined using a given function $g(\cdot)$ as follows :

$$\hat{x}^{d_i} = g(\frac{\alpha^{d_i}}{\alpha^{d_i} + \alpha^{d_j}} x^{d_i}, \frac{\alpha^{d_j}}{\alpha^{d_i} + \alpha^{d_j}} y^{d_i}) \qquad (3)$$

where $\hat{x}^{d_i}$ are the combined features belonging to domain d_i, y^{d_i} is the feature projected to the d_i domain, α^{d_i} and α^{d_j} are the feature attention scores and $g(\cdot)$ is a summation.

Finally, $\hat{x}^{d_i}$ is used to weight the input feature vector x^{d_i}. To this end, a gating over these features is performed by applying a sigmoid function to the combined features.

3. Pyramid Feature Combination

Inspired by the satisfactory performance of feature pyramid networks for object recognition [9], a multi-modal Pyramid Feature Combination method is proposed in this section. This technique combines features from different domains at multiple hierarchical levels. Specifically, a bottom-up pathway refines action features with the multi-modal information from other domains, while the top-down pathway uses these refined features to enrich the ones at lower pyramid levels.

Bottom-up pathway. The bottom-up pathway is the feedforward computation of the combination of multi-modal features by pairs, which computes a feature hierarchy consisting of feature vectors at several scales. At each scale a pair of feature vectors are combined using the MFF presented in Section 2.3. The resulting feature vector after this combination has the same dimension as each input vector. Thus, as in the case of FPN for object recognition, the feature dimensionality is reduced at each level of the pyramid. However, instead of reducing the features by a fixed factor of 2 at each scale as in RPN, they are reduced by a factor $(L-l)/(L-l+1)$ where l is the level of the pyramid and L is the total number of levels.

The gated feature vector at each level is chosen as the reference feature which is further leveraged to create the pyramid. This is the most suitable choice for this task as it gathers the information of the multi-modal input.

The feature vector at the top of this bottom-up path represents the refined feature for human action recognition using all the other features. On the other hand, the bottom of this path is associated with the initial multi-modal features without any combination. The order in which features are combined is further discussed in Section 4.

Top-down pathway and lateral connections. The top-down pathway hallucinates features by upsampling feature vectors from higher pyramid levels which are coarser, but semantically stronger for the task of human action recognition. These features are then leveraged with features from the bottom-up pathway via lateral connections. Each lateral connection combines feature vectors of the same dimension from both pathways (bottom-up and top-down).

The dimension of the feature vector from the higher level is increased by a factor of $(L-l+2)/(L-l+1)$ using a fully connected layer with ReLU activation function, where l is the level of the pyramid and L is the total number of levels. Then, the resulting feature is added to the corresponding bottom-up feature vector. This process is iterated until the number of initial feature vectors are generated.

4. Experiments

In this section, the experiments to assess and prove the robustness of the proposed methods are described in detail.

All the experiments have been performed using a subset of the Moments In Time dataset [11]. The MIT dataset was developed to improve trimmed action recognition tasks and contains over 1 million 3-second clips annotated with their actions. The motivation behind the reduction of MIT is to focus into Human Action Recognition. Thus, only samples annotated with a human action have been taken to conduct the presented experiments. The pipeline used for human action detection uses a Yolov2 [12] trained with MSCOCO. The Yolov2 network has been chosen over a Faster R-CNN for computational performance reasons. The resulting subset is composed of 129 classes, with around 130K videos for train and 4K for validation. All the following results are reported on the validation subset.

4.1. Single feature experiments

In order to assess the proposed techniques, pretrained features for audio, scene, objects and actions have been used in all the experiments as shown in Section 2.1. The networks used for feature extraction are a ResNet50 for objects and scenes, trained on Imagenet [3] and Places365 [14] respectively. Then, an *I3D ConvNet* [2] trained on the Kinetics dataset [8] and finally, a VGG-like network for audio feature extraction [6], trained on the AudioSet dataset [4].

The number of frames used for feature extraction has been set to 64 for action features and 5 for scene and objects. The temporal dimension is reduced following the approach described in Section 2.2.

The single feature experiment consists on using the features extracted from each network for human action recognition. For fair comparison, the feature dimensionality is set to 1024 by adding a fully-connected layer before classification. The results of this experiment are shown in Table 1 (Left column). As it can be observed, action features work better than the others because they have been extracted from a network trained on a very similar task to the human action recognition task with the Moments In Time subset that is performed in this work.

4.2. Pyramid experiments

In this section, an experiment designed to evaluate the effectiveness of the pyramid combination approach is presented. The size of the pyramid is increased by adding more feature domains until all the extracted features are combined, as it is shown in Figure 3.

In order to perform the final classification, features extracted from all the levels of the pyramid are combined. To this end, a bottleneck layer reduces the feature dimensions to 1024, followed by sum pooling and a classification layer.

The results obtained for different pyramid sizes are shown in Table 2. As it can be observed in this table, increasing the number of feature domains results in richer features generated by the pyramid. This is reflected by obtaining better accuracy, reaching a top-1 accuracy of **35.43%** when all the extracted features are combined.

Feature type	Base feature	Enriched feature
Actions	24.97%	30.07%
Scene	17.43%	30.64%
Objects	15.53%	29.2%
Audio	15.63%	28.36%

Table 1. Results of individual features for action recognition in the Moments In Time subset. Corresponding to the individual base feature accuracy (left) and enriched feature accuracy (right).

Pyramid combinations	Accuracy
Actions	24.97
Actions**+Scenes**	25.66
Actions+Scenes**+Audio**	33.36
Actions+Scenes+Audio**+Objects**	**35.43**

Table 2. Results of the multi-modal feature pyramid combination for action recognition in the Moments In Time subset.

Moreover, the richness of features from the different domains extracted by the pyramid is assessed in the second column of Table 1. This table shows the results of the single feature experiment presented in Section 4.1 using the features of each domain extracted at the bottom of the hierarchy (bottom-up combined with top-down). As it can be observed, for each type of feature, the action recognition accuracy is significantly increased.

5. Conclusions

Accurate human action recognition is one of the most challenging tasks in computer vision. While initially addressed by the research community through the usage of narrow image features (which can be enough to distinguish certain classes), this paper analyzes the addition of task specific features (scene understanding, object detection and acoustic event detection) in order to achieve more fine-grained action understanding.

This work comprises a two-fold contribution: 1) a novel fusion block that performs domain-specific attention and combines pairs of features from different domains (audio and video) and 2) a pyramid approach that uses the previous fusion block at different levels of a hierarchy, resulting in a system that refines the input task-specific features (bottom-up pathway) and enriches features from other domains (top-down pathway) for human action recognition.

Both contributions demonstrate that multi-modal features can robustly improve human action recognition if efficiently combined. The experiments presented in Section 4 show a significant accuracy increase (from 24.97% using action features, to **35.43%** with the full pyramid) on a subset of Moments in Time dataset (human actions).

References

[1] A. Bhoi. Spatio-temporal action recognition: A survey. *CoRR*, abs/1901.09403, 2019.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[5] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. A better baseline for AVA. *CoRR*, abs/1807.10066, 2018.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[7] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 494–507, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[9] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[10] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[11] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. M. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018.

[12] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[13] Z. Zhao and M. Larson. From volcano to toyshop: Adaptive discriminative region discovery for scene recognition. *CoRR*, abs/1807.08624, 2018.

[14] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.