Temporal Regions for Activity Recognition

João Paulo Aires[1], Juarez Monteiro[1], Roger Granada[1], Felipe Meneguzzi[2], and Rodrigo C. Barros[2]

Faculdade de Informática
Pontifícia Universidade Católica do Rio Grande do Sul
Av. Ipiranga, 6681, 90619-900, Porto Alegre, RS, Brazil
[1] Email: {joao.aires.001, juarez.santos, roger.granada}@acad.pucrs.br
[2] Email: {rodrigo.barros, felipe.meneguzzi}@pucrs.br

Abstract. Recognizing activities in videos is an important task for humans, since it helps the identification of different types of interactions with other agents. To perform such task, we need an approach that is able to process the frames of a video and extract enough information in order to determine the activity. When dealing with activity recognition we also have to consider the temporal aspect of videos since activities tend to occur through the frames. In this work, we propose an approach to obtain temporal information from a video by dividing its frames into regions. Thus, instead of classifying an activity using only the information from each image frame, we extract and merge the information from several regions of the video in order to obtain its temporal aspect. To make a composition of different parts of the video, we take one frame of each region and either concatenate or take the mean of their features. For example, consider a video divided into three regions and each frame containing ten features, the resulting vector of a concatenation will contain thirty features, while the resulting vector of the mean will contain ten features. Our pipeline includes pre-processing, which consists of resizing images to a fixed resolution of 256×256; Convolutional Neural Networks, which extract features from the activity in each frame; region divisions, which divides each sequence of frames of a video into n regions of the same size; and classification, where we apply a Support Vector Machine (SVM) on the features from the concatenation or mean phase in order to predict the activity. Experiments are performed using The DogCentric Activity dataset [1] that contains videos with 10 different activities performed by 4 dogs, showing that our approach can improve the activity recognition task. We test our approach using two networks *AlexNet* and *GoogLeNet*, increasing up to 10% of precision when using regions to classify activities.

Keywords: Neural Networks, Convolutional Neural Networks, Activity Recognition

1. Iwashita, Y., Takamine, A., Kurazume, R., Ryoo, M.S.: First-person animal activity recognition from egocentric videos. In: ICPR' 14 (2014)