2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning Supplementary Material

Diogo C. Luvizon[1], David Picard[1,2], Hedi Tabia[1]
[1]ETIS UMR 8051, Paris Seine University, ENSEA, CNRS, F-95000, Cergy, France
[2]Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, LIP6, F-75005 Paris, France
{diogo.luvizon, picard, hedi.tabia}@ensea.fr

Appendix A: Network architecture

In our implementation of the proposed approach, we divided the network architecture into four parts: the *multitask stem*, the *pose estimation model*, the *pose recognition model*, and the *appearance recognition model*. We use depth-wise separable convolutions as depicted in Figure 1, batch normalization and ReLu activation. The architecture of the multitask stem is detailed in Figure 2. Each pose estimation prediction block is implemented as a multi-resolution CNN, as presented in Figure 3. We use N_d = 16 heat maps for depth predictions. The CNN architecture for action recognition is detailed in Figure 4.

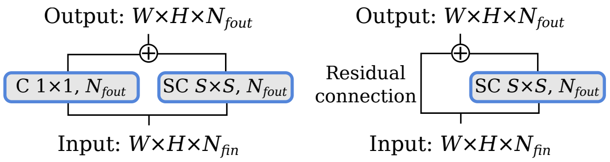
Output: $W{\times}H{\times}N_{fout}$
C 1×1, N_{fout}
SC $S{\times}S$, N_{fout}
Input: $W{\times}H{\times}N_{fin}$
Output: $W{\times}H{\times}N_{fout}$
Residual connection
SC $S{\times}S$, N_{fout}
Input: $W{\times}H{\times}N_{fin}$

Figure 1. Separable residual module (SR) based on depth-wise separable convolutions (SC) for $N_{fin} \neq N_{fout}$ (left), and $N_{fin} = N_{fout}$ (right), where N_{fin} and N_{fout} are the input and output features size, $W \times H$ is the feature map resolution, and $S \times S$ is the size of the filters, usually 3×3 or 5×5. C: Simple 2D convolution.

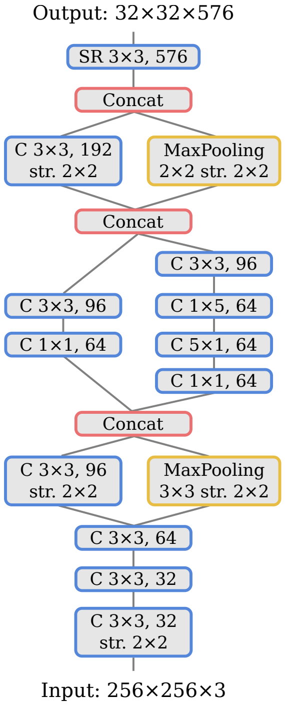

Figure 2. Shared network (entry flow) based on Inception-V4. C: Convolution, SR: Separable residual module.

Appendix B: Training parameters

In order to merge different datasets, we convert the poses to a common layout, with a fixed number of joints equal to the dataset with more joints. For example, when merging the datasets Human3.6M and MPII, we use all the 17 joints in the first dataset and include one joint on MPII. All the included joints have an invalid value that is not taken into account in the loss function. Additionally, we use and alternated human pose layout, similar to the layout from the Penn Action dataset, which experimentally lead to better scores on action recognition.

We optimize the pose regression part using the RMSprop optimizer with initial learning rate of 0.001, which is reduced by a factor of 0.2 when validation score plateaus, and batches of 24 images. For the action recognition task, we train both pose and appearance models simultaneously using a pre-trained pose estimation model with weights initially frozen. In that case, we use a classical SGD optimizer with Nesterov momentum of 0.98 and initial learning rate of 0.0002, reduced by a factor of 0.2 when validation plateaus, and batches of 2 video clips. When validation accuracy stagnates, we divide the final learning rate by 10 and fine tune the full network for more 5 epochs. When reporting only pose estimation scores, we use eight prediction blocks ($K = 8$), and for action recognition, we use four prediction blocks ($K = 4$). For all experiments, we use cropped

Table 1. Our results on averaged joint error on reconstructed poses for 3D pose estimation on Human3.6 considering single dataset training (Human3.6M only) and mixed data (Human3.6M + MPII). SC: Single-crop, MC: Multi-crop.

Methods	Direction	Discuss	Eat	Greet	Phone	Posing	Purchase	Sitting	Sit Down	Smoke	Photo	Wait	Walk	Walk Dog	Walk Pair	Average
Human3.6 only - SC	64.1	66.3	59.4	61.9	64.4	59.6	66.1	78.4	102.1	67.4	77.8	59.3	51.5	69.7	60.1	67.3
Human3.6 only - MC	61.7	63.5	56.1	60.1	60.0	57.6	64.6	75.1	95.4	63.4	73.3	57.0	48.2	66.8	55.1	63.8
Human3.6 + MPII - SC	51.5	53.4	49.0	52.5	53.9	50.3	54.4	63.6	73.5	55.3	61.9	50.1	46.0	60.2	51.0	55.1
Human3.6 + MPII - MC	49.2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7	60.3	48.9	44.4	57.9	48.9	53.2

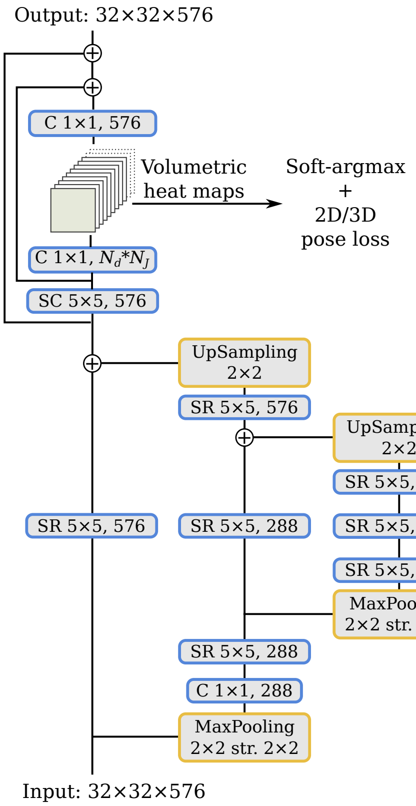

Figure 3. Prediction block for pose estimation, where N_d is the number of depth heat maps per joint and N_J is the number of body joints. C: Convolution, SR: Separable residual module.

RGB images of size 256 × 256. We augment the training data by performing random rotations from $-45°$ to $+45°$, scaling from 0.7 to 1.3, vertical and horizontal translations respectively from -40 to $+40$ pixels, video subsampling by a factor from 1 to 3, and random horizontal flipping.

Appendix C: Additional experiments

In order to show the contribution of multiple datasets in training, we show in Table 1 additional results on 3D pose estimation using Human3.6M only and Human3.6M + MPII datasets for training.

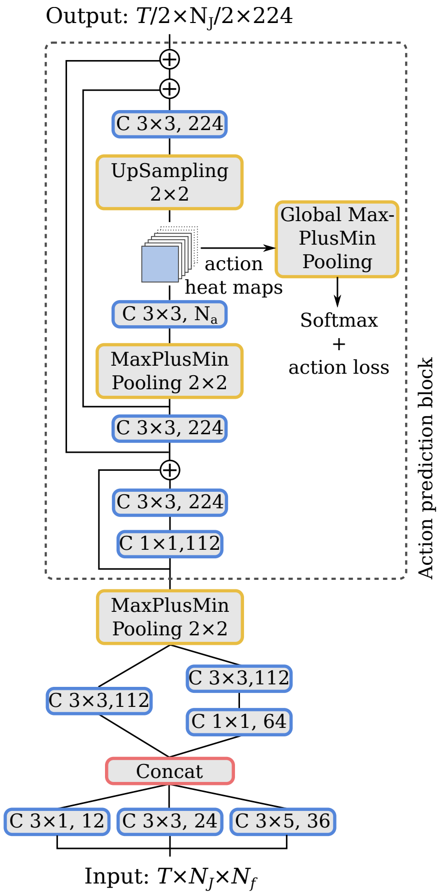

Figure 4. Network architecture for action recognition. The action prediction blocks can be repeated K times. The same architecture is used for pose and appearance recognition, except that for pose, each convolution uses half the number of features showed here. T corresponds the number of frames and N_a is the number of actions.