# Heterogeneous Non-Local Fusion for Multimodal Activity Recognition

Petr Byvshev
petr.bysvhev@aalto.fi
Aalto University
Espoo, Finland

Pascal Mettes
University of Amsterdam
Amsterdam, Netherlands

Yu Xiao
Aalto University
Espoo, Finland

## ABSTRACT

In this work, we investigate activity recognition using multimodal inputs from heterogeneous sensors. Activity recognition is commonly tackled from a single-modal perspective using videos. In case multiple signals are used, they come from the same homogeneous modality, e.g. in the case of color and optical flow. Here, we propose an activity network that fuses multimodal inputs coming from completely different and heterogeneous sensors. We frame such a heterogeneous fusion as a non-local operation. The observation is that in a non-local operation, only the channel dimensions need to match. In the network, heterogeneous inputs are fused, while maintaining the shapes and dimensionalities that fit each input. We outline both asymmetric fusion, where one modality serves to enforce the other, and symmetric fusion variants. To further promote research into multimodal activity recognition, we introduce GloVid, a first-person activity dataset captured with video recordings and smart glove sensor readings. Experiments on GloVid show the potential of heterogeneous non-local fusion for activity recognition, outperforming individual modalities and standard fusion techniques.

## CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

## KEYWORDS

datasets, activity recognition, heterogenous modalities

**ACM Reference Format:**
Petr Byvshev, Pascal Mettes, and Yu Xiao. 2020. Heterogeneous Non-Local Fusion for Multimodal Activity Recognition. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3372278.3390675


*ICMR '20, June 8–11, 2020, Dublin, Ireland*

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390675

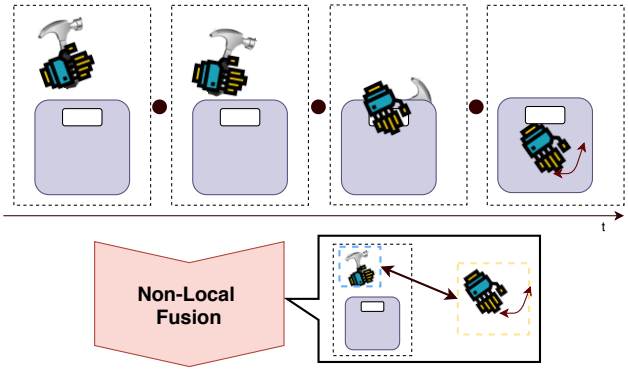


**Figure 1: Non-Local Fusion.** Recognizing an activity from a video clip with complex dynamics states a challenging problem. Combining video inputs with a sensor (here, a smart-glove) in a non-local manner allows us to recognize activities, even when they are not visible on camera.

## 1 INTRODUCTION

This paper proposes an approach that can recognize activities captured by multiple modalities from vastly different sensors. Activity recognition from videos has gained a lot of traction in recent years, due to advances in deep networks designed for videos [7, 17, 18, 58, 59, 62]. These advances are fueled by the introduction of new large-scale video datasets such as Moments in Time [42], Kinetics [32], and ActivityNet [28]. The ability to automatically recognize activities and events is an important component in video surveillance [1, 43, 46], virtual/augmented reality [5, 24], and human-robot interaction [3, 25, 41]. Here, we aim to improve activity recognition by incorporating information that is missed or inherently invisible purely from videos.

Often human actions look like a magic trick - a sleight of hand, for a machine vision method. Certain hand movements or operated objects that define an activity can be occluded or misinterpreted for video inputs, see Figure 1. A fixed camera might miss an important visual clue due to occlusions or insufficient resolution, and a body-mounted camera, having a limited field of view, might not capture an activity region of interest. On the other hand, sensor-based activity recognition relies on time sequence data collected by various wearable sensors while missing the necessary context information about the surrounding environment may successfully recognize the differences between certain body dynamics [9]. The information extracted from sensors can

help reliably differentiate between activities with pronounced dynamics, where the context and surroundings do not define the action category. The challenge in combining videos and wearable sensors for activity recognition is that they are heterogeneous in two aspects; the type of information coming from the inputs is different and the modalities do not match in their spatio-temporal shapes and topologies.

In this paper, we propose heterogeneous non-local fusion, a fusion component that allows for end-to-end learning of activities in a unified multimodal network. We seek to learn complementary information across modalities, without the need to match the shapes and dimensions of the modalities. We do so by generalizing non-local neural networks, originally designed for self-attention [59]. The key behind our fusion is that non-local operations can be performed on different inputs; the spatio-temporal dimensions do not even need to match. Only the channel dimensions need to be of equal size. The result is a network that combines modalities, while abiding to their heterogeneous nature. We outline both asymmetric and symmetric variants of the fusion. To promote research into heterogeneous multimodal activity recognition, a new dataset is proposed: GloVid. The GloVid dataset contains samples captured by both mounted cameras and smart gloves in a maintenance setting. As illustrated in Figure 1, smart gloves provide information hidden from video cameras and vice versa. Experiments on the new dataset show that our fusion provides a direct boost over the individual modalities and outperform common fusion techniques. The code, as well as preprocessed multimodal features and labels for samples in GloVid will be made accessible upon acceptance.

## 2 RELATED WORK

### 2.1 Video-based Activity Recognition

Traditionally, activity recognition from video sequences was tackled through crafted spatio-temporal features [15, 36, 38] aggregated over videos [34, 57, 60]. Within the last years, video recognition with deep spatio-temporal networks has become the dominant research direction. Well-known network architectures include two-stream networks [20, 48], long-term temporal convolutions [52], and multiplier networks that motion and appearance pathways through multiplicative gating [19]. Their main idea is to learn representations from both RGB and optical flow streams to learn effective space and time features. More recently, 3D convolutional networks have shown to be state-of-the-art for video-based activity recognition. Inflated 3D networks (I3D) use expanded 2D filters and pooling layers trained on image classification [7]. 3D residual connection networks (Resnet3D) proved to be adequate in case of sufficient amount of data as they are prone to over-fitting [26]. 3D convolutional networks can also be factorized into (2+1)D networks for parameter efficiency [51]. All such approaches are trained on large-scale datasets such as ActivityNet [28], Kinetics [32], Moments in Time [42], and Sports-1M [30]. Such datasets provide hundreds of activity classes with hundreds/thousands of samples per class. As an extension to current approaches, the non-local network has gained popularity as a way to deal with the non-local nature and long-range dependencies in videos [59]. A non-local block follows a convolutional operation and learns relations between different positions (in space, time, or space-time) of a video representation. In [10], cross-modal attention is performed on homogeneous modalities, namely RGB and optical flow, to learn non-local relations. In our work, we rely on state-of-the-art 3D convolutional networks to represent video modalities and we are inspired by the non-local block, which we generalize to operate on multiple inputs simultaneously, even when their shapes and dimensions do not directly correspond.

### 2.2 Sensor-based Activity Recognition

Wearable sensors such as contact sensors, RFIDs (Radio Frequency Identification), IMU (Inertial Measurement Unit) composed of accelerometers, gyroscopes, magnetometers, and microphones have been used to detect events and activities [23, 35, 45, 61, 63]. Such sensors are attached to different body parts depending on their functions and application domain. Hand mounted sensors such as smart rings, gloves, watches, bands, and smartphones have recently also gained traction for activity understanding [4, 39, 49]. Initial work focused on recognition from traditional machine learning approaches, such as decision trees, support vector machines and random forests [13, 16, 21, 40, 50]. For tasks such as tracking in sports activities, intensity and count repetition estimation even suffices [2, 56]. Nonetheless, deep learning solutions have achieved state-of-the-art performance in the context of activity recognition, where training data is represented by temporal sensor recordings. For activity classification (sensor readings classification) one-dimensional convolutions and long short-term memory units have been used as basic building blocks [11, 29, 64, 65]. Such networks learn local correlations with 1D convolutions and long-term dependencies with the LSTM module from raw sensor input.

Combining sensor inputs within a single deep network is a less commonly tackled problem. Sensory combinations have been studied using late convolutional fusion followed by fully-connected layers [31], using convolutional LSTMs on top of sensory-independent representations [44], or through multi-objective regularization on different embedding combinations for sensory-specific networks [54]. Such approaches focus on fusing low-dimensional sensors, ignoring high-dimensional signals such as videos. Multimodal fusion for activity recognition with videos is generally homogeneous, e.g. fusing color and optical flow streams [10, 47]. In heterogeneous settings, the non-video sensor generally has the same spatial and temporal dimensionalities, such as with RGB-D cameras, allowing for a fusion along those dimensions [8, 12]. When shapes and topologies also do not match, a late fusion of final scores from independently trained single-modal networks is performed [33, 55]. In this work, we seek to fuse heterogeneous inputs from videos and smart gloves, which have non-matching spatio-temporal topologies. Rather than fusing the inputs after independent training, we propose a

fusion mechanism that integrates both signals within a single network. Fermüller et al. [22] also employ videos and hand sensors, but do so to predict finger forces from videos. We focus on fusing both signals for activity recognition.

# 3 NON-LOCAL FUSION NETWORK

## 3.1 Problem Formulation

For the problem of multimodal activity recognition, we aim to learn a function $\phi(\mu_1, \mu_2) \mapsto c$, where $\mu_1$ and $\mu_2$ denote the heterogeneous sensory inputs of a sample and $c \in C$ denotes the predicted class label from a set $C = \{1, ..., |C|\}$. A challenge in this problem setting is that both the input dimensionality and shape of the modalities vary. The number of dimensions typically varies from 1-dimensional (a temporal sequence of sensor readings) to 3-dimensional (space and time dimensions of a video). Moreover, the temporal and spatial extents of the modalities may not be aligned. To tackle this problem, we propose to formalize the function $\phi(\cdot, \cdot)$ as a network that is able to fuse the video modalities in spite of their non-matching shapes and dimensions.

To train our proposed network, we are given a training set of $N$ double-modality tuples $\{(\mu_1^{(i)}, \mu_2^{(i)}, c^{(i)})\}_{i=1}^{N}$. Our network consists of two branches that perform representation learning of the individual modalities: $\psi(\mu_1, \mu_2) \mapsto (x, y)$, followed by a novel heterogeneous non-local fusion block to combine the representations. The block is followed by a few final layers, resulting in $|C|$-dimensional output representation. The output representation is fed to a soft-max to obtain activity probabilities, which are compared to the ground truth activity labels and optimized using the cross-entropy loss. The key to getting an effective network is in the fusion of the heterogeneous video representations.

## 3.2 Non-Local fusion

**Non-local block.** For our heterogeneous fusion, we take inspiration from the non-local block, which was originally designed for the purpose of self-attention in deep networks [59]. The idea behind the non-local block is to compute relations between different space-time locations of the same instance. The standard non-local operation is defined as follows:

$$z_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{1}$$

where $x$ is an input signal, $i$ is the index of an output position, $j$ enumerates all other possible positions, and $f$ computes a scalar relation between $i$ and all $j$. Function $g$ computes a representation of the input signal at position $j$ and $C(x)$ is a normalization factor. In this formulation, each spatio-temporal location is enforced by representations from all other locations. Here, we seek to generalize the idea of the non-local operation to take in multiple inputs.

**Heterogeneous non-local fusion.** The idea behind our heterogeneous non-local fusion is to generalize the non-local block in two aspects. First, we change the perspective from self-attention to cross-modal attention by using two multi-modal inputs, rather than the same input twice. Second, we show how to utilize the non-local operation in such a way that the spatio-temporal dimensions do not need to match in order to perform the fusion. At a high-level, this operation

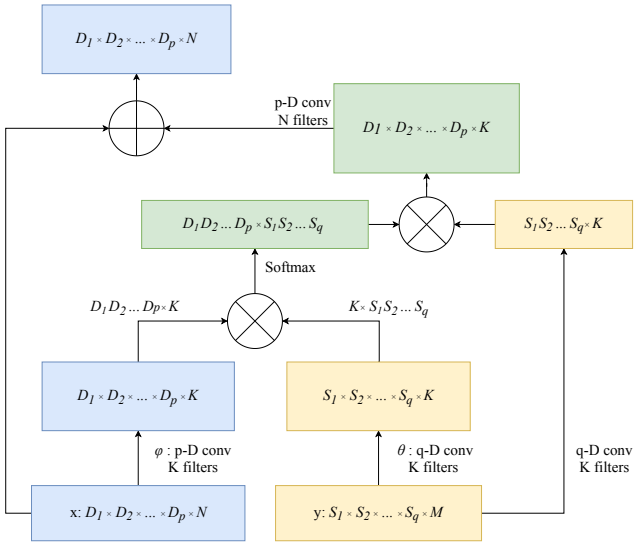


**Figure 2: The Non-Local Fusion block.** x and y - two modality feature tensors with varied shape. At each stage we show the shape of a feature with the last value being the number of channels (in $D_1 \times D_2 \times ... \times D_p \times \underline{\mathrm{N}}$ is the number of channels). $\bigotimes$ and $\bigoplus$ denote matrix multiplication and tensor-sum.

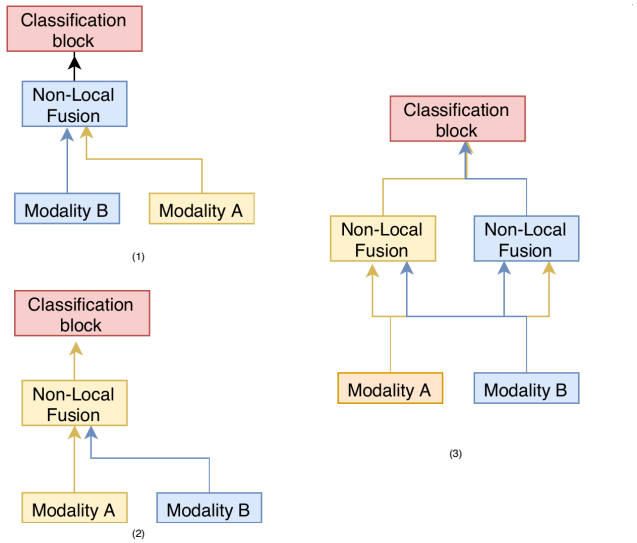


**Figure 3: 3 variants of the Non-Local Fusion.** (1) and (2) - the fusion is built upon 1st or 2nd modality. (3) Both modalities are augmented with the non-local fusion.

changes to:

$$z_i = \frac{1}{C(x_i, y)} \sum_{\forall j} f(x_i, y_j) g(y_j), \quad (2)$$

$$f(x_i, y_j) = e^{\theta(x_i)^T \phi(y_j)}, \quad (3)$$

$$C(x_i, y) = \sum_{\forall j} f(x_i, y_j). \quad (4)$$

In this formulation, $i$ is the dimensional index of the first modality ($x$) and $j$ of the second modality ($y$). In other words, modality $y$ enforces the features of modality $x$. This is done through function $f$, which should now model the relationship between $x$ and $y$ and $g$ that maps $y$ into the same embedding space. Our observation for function $f$ is that the two modalities only need to match across the channel dimensions, all other dimensions can be different. To that end, we model $f$ as a network block, where $x$ is a tensor of shape $D_1 \times D_2 \times ... \times D_p \times N$, where $p$ is the number of dimensions, $\{D_k\}_{k=1..p}$ dimensions' sizes, and $N$ is the number of feature maps (channels). In similar fashion, $y$ is a tensor of shape $S_1 \times S_2 \times ... \times S_q \times M$, where $p$ and $q$ do not necessary match.

Our non-local fusion block is shown in Figure 2. First, functions $\theta(.)$ and $\phi(.)$ perform convolution across all non-channel dimensions for $x$ and $y$ respectively, both using $K$ feature maps. This ensures an alignment across the channel dimension. Consequently, we perform a matrix product along the channel dimension of both output tensors, followed by a soft-max to obtain a form of multimodal attention. Afterwards, we perform a matrix product over the non-channel dimensions of $y$ to factor out those dimensions, resulting in a transformed version $z$ of input $x$ by the hand of $y$. Then, we perform a convolutional operation $h$ over the non-channel dimensions of the transformed version of $x$ to revert it back to the original number of $N$ channels. Lastly, we obtain the block output by adding a residual connection to the input $x$ itself:

$$\hat{z} = h(z(x, y)) + x. \quad (5)$$

The method can be extended to more then two modalities by applying the non-local operation to all the combinations of the inputs:

$$\hat{z} = h^1(z^1(x, y^1)) + h^2(z^2(x, y^2)) + ... + h^r(z^r(x, y^r)) + x, \quad (6)$$

where $h^1, h^2, ...h^r$ compute relations between the primal modality $x$ and all the others $y^1, y^2, ...y^r$.

**Asymmetric and symmetric fusion.** By design of our non-local fusion, one modality is transformed using the representations of the other modality. To investigate which fusion ordering is preferred for activity recognition, we outline three variants, as shown in Figure 3. The first two directly follow from the non-local fusion itself; one modality is the main modality, the other serves as transformer. Next to these asymmetric fusions, we also investigate a symmetric fusion, where the non-local fusion is done with both modalities as main, followed by a classification block. Since the symmetric fusion keeps the corresponding shapes of input tensors we need an extra fusion block for the final classification. We test concatenation of the corresponding non-local fusion outputs followed by a fully-connected layer.

# 4 EXPERIMENTAL SETUP

## 4.1 Dataset

**The GloVid dataset.** To promote research into activity recognition from multiple heterogeneous sensors, we introduce a novel dataset. The dataset consists of eight activities that heavily involve hand dynamics, centered around elevator maintenance. The context and background are similar for all activities, which makes recognition challenging since static appearance is not sufficient.

For multimodality, the dataset is constructed by capturing both first-person video information and sensor readings from gloves. The dataset studies a typical workflow procedure, with eight activities that commonly occur during such a workflow. The activities are: *neutral activity, pressing a button, unplugging the cables, plugging the cables, disassembling a button, assembling a button, using a screwdriver,* and *using a hummer*. A visual example for each activity is shown in Figure 4.

**Dataset details.** The dataset was recorded with help of 19 subjects, each performing a maintenance procedure for an elevator panel. Each subject takes 16 to 20 minutes to complete the list of steps, sometimes taking breaks in-between, which are assessed as neutral activity. For data collection we use an RGB camera and a pair of Captogloves [6]. Each glove is equipped with flex sensors for the five fingers, an accelerometer, a gyroscope, a magnetometer, and a pressure sensor for the thumb. The videos are collected with a chest mounted Gopro camera (Gopro HERO7, 1080x1920, 60fps), which provides an egocentric view of the workflow. Before performing the task, each participant clapped their hands indicating the start of the recording for the camera and the smart-gloves.

The GloVid dataset consists of 3800 video-glove samples, 200 for each subject. Original 60fps videos were downsampled to 15fps. Fixed-sized chunks of 79 frames are extracted to be used as individual examples, amounting to roughly 5.3 seconds. Smart-glove sensory reading are extracted to match the duration of video chunks, corresponding to 134 readings at 25Hz. Throughout the experiments, we use the flex sensors in each finger (5D vector), the accelerometer (3D vector) and the pressure sensor (1D vector) for each hand, resulting in a 18-dimensional vector. This way each sample in the GloVid dataset consists of a pair - $(79 \times 224 \times 224 \times 3$ and $134 \times 18)$, that represents a video chunk and smart glove readings and a one-hot encoded activity class.

## 4.2 Network Architectures

To obtain the video features we use the I3D network pretrained on the Kinetics dataset [7]. Each video sample is represented by the feature of shape $10 \times 7 \times 7 \times 832$ extracted from the *'Mixed_5b'* layer of the I3D network. The glove


**Figure 4: A visual example for each of the eight activities in the new GloVid dataset.** The examples show the potential of recognition from multiple modalities, since the activities are determined by spatio-temporal hand dynamics and visual context.

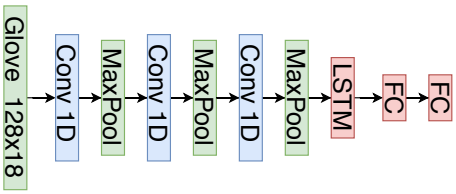


**Figure 5: Architecture of the glove network.** Smart gloves generate 1D sensory signals, which are used as input for a network which performs several 1D convolutions, followed by an LSTM and fully-connected layer to obtain video-level representations.

subnetwork is trained from scratch on raw sensor readings and follows the architecture similar to [65]: we use three 1D convolution+max-polling operations followed by a bidirectional LSTM block, a fully-connected layer, and a final soft-max classification layer. Figure 5 shows an overview of the glove network architecture. For the feature extraction we use the output of the LSTM layer of shape: $28 \times 64$, where first number represents the time steps, and second - dimensionality of the feature space.

### 4.3 Training and Evaluation Metrics

The Glovid dataset consists of recordings of 19 subjects. We test all the models using cross-subject evaluation: each model is trained 19 times, each time 16 subjects are in the training set 2 in the validation set and the remaining one serves as a test set. We use standard cross-entropy loss and RMSProp optimization (learning rate: 0.001, exponentially weighted average: 0.9) with batch size: 8. We terminate the training if validation set accuracy does not improve for 20 epochs. The mean accuracy across the 19 subject-specific initiations is calculated for each model.

## 5 EXPERIMENTAL RESULTS

### 5.1 Comparative evaluation

First, we perform evaluations to (i) compare our fusion to individual modalities, (ii) compare the three proposed fusion variants, and (iii) compare our fusion with standard fusion approaches.

**Comparison to individual modalities.** In the first evaluation, we compare our proposed non-local fusion to the performance of the individual modalities, to gain insight into the effect of fusion. We compare our non-local fusion to three individual modalities. The first two are the modalities that are used in the fusion, namely the smart gloves and the video deep network. We also compare to a 2D ResNet50 baseline pretrained on Imagenet [14], which combines the 2D features from the last convolution layer with 3D global pooling for classification. This baseline serves to show whether the activities in the dataset can be recognized from static appearance only.

We show the result in Table 1. Individually, the modalities obtain an accuracy of 46.4% (2D ResNet50), 68.1% (Smart gloves), and 76.3% (Video I3D). The video-based model works best, which can partially be attributed to the scale and pretraining in the I3D network. The 2D ResNet50

| | **Accuracy** (%) | **Subject SD** |
|---|---|---|
| **Single modality** | | |
| 2D ResNet50 [27] | 46.4 | 8.5 |
| Smart gloves | 68.1 | 10.5 |
| Videos (I3D [7]) | 76.3 | 7.9 |
| **Non-local fusion** | | |
| Asymmetric (glove-led) | 80.1 | 8.2 |
| Asymmetric (video-led) | **80.8** | 7.7 |
| Symmetric | 80.2 | 8.0 |

**Table 1: Evaluation of modalities on GloVid.** All our non-local fusion variants boost the performance over videos and smart gloves individually. The 2D ResNet50 baseline can not compete due to lack of temporal information. The asymmetric fusion where smart gloves enforce the features of the main video modality performs slightly better than the other variants.

| | **Accuracy** (%) | **Subject SD** |
|---|---|---|
| Average score fusion | 78.9 | 8.2 |
| Late feature concatenation | 79.3 | 7.9 |
| This paper | **80.8** | 7.7 |

**Table 2: Comparison of fusions on GloVid.** Both standard fusion baselines do not perform as well as our approach, highlighting the potential of our fusion for mulitmodal activity recognition.

scores much lower, showing that this dataset can not easily be solved with static representations, temporal understanding is vital. We fuse the gloves and videos, resulting in an accuracy of 80.8%, an improvement of 12.7 percent point (p.p.) compared to smart gloves and 4.5 p.p. compared to videos. These results clearly indicate that non-local fusion provides a direct boost to the recognition performance, as the complementary information from the heterogeneous sensors is elevated. As we train all the models 19 times to get a subject specific performance we compute the cross-subject standard deviation (SD). The SD values are relatively high (7.7-10.5) which shows that cross-subject differences, such as camera position, hand size and preferred body pose have a strong effect on the performance. Fusion helps to reduce this effect.

**Comparison of fusion variants.** The non-local fusion block is a non-symmetrical operation, it outputs activations of the first modality re-weighted by the second. These leads to 3 possible fusion setups: video-led, gloves-led, and a symmetrical fusion. Table 1 shows that all three variants outperform the individual modalities. The fusion led by the video modality and augmented with smart gloves readings performs slightly better and is therefore chosen for the future evaluation. This outcome follows the results for the individual modalities, where the video-based approach performed best.

**Comparison to standard fusion.** Finally, we compare our approach to two conventional fusion approaches, namely average score fusion and late feature concatenation. For average score fusion, we train independent networks for both modalities and average the activity probability scores for test videos. For late feature concatenation, we flatten and concatenate features from the *Mixed_5b* layer of the video network and LSTM layer of the glove network. On top of the concatenation, we add a fully-connected soft-max classification layer. Table 2 shows that the proposed fusion method outperforms the baselines achieving 80.8% accuracy, compared to 78.9% and 79.3%. This result shows that for the same inputs and network architectures, our approach is capable of learning more complementary representations.

## 5.2 Quantitative analysis

To gain more insight into the workings of our approach, we perform three quantitative evaluations, focused on per activity effect of the fusion, the effect of the embedding layer size, and the effect of hand visibility.

**Per activity improvement of our fusion.** Figure 6 depicts the improvement of the fusion model over the glove and video models with respect to activity categories. We find that for all the categories the non-local fusion is equal or better than the unimodal solutions. For the *neutral* activity, the low improvement over videos and high improvement over gloves can be explained by the non-information nature of the glove readings for recognizing neutral behaviour. The high improvement of the *using screwdriver* category reasons that both modalities require extra information that is learned by the proposed fusion.

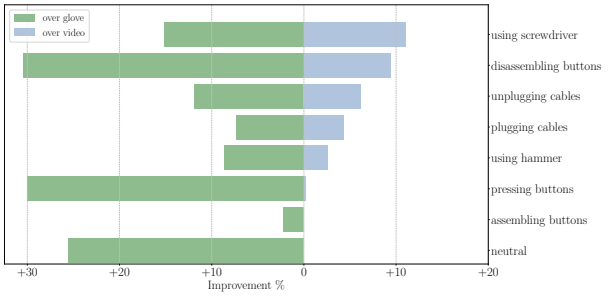


**Figure 6: Per activity improvement of our fusion.** The activities are sorted by their improvement over the video modality. The improvement over gloves are bigger, while no activity is negatively impacted by the fusion.

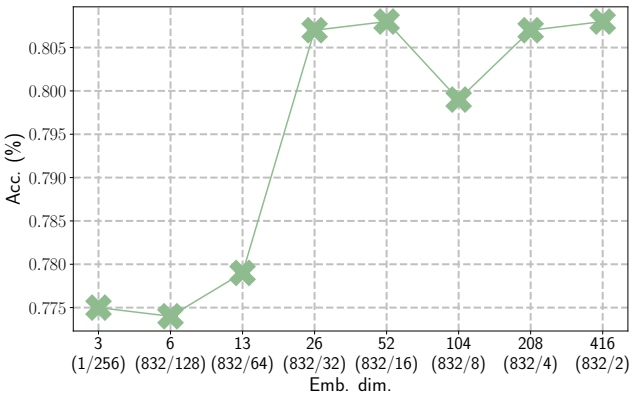


**Figure 7: Effect of embedding layer size.** The x-axis is in logarithmic scale and depicts the fraction of the input number of channels (a fraction of the 832 input dimensions). We find that the more dimensions in the embedding layer, the better performance, although a low dimensionality can already obtain high scores.

**Effect of embedding layer size.** We evaluate the model performance by varying the embedding space dimensionality $K$. Typically in a non-local block $K$ is chosen to be half of the number of channels in the input (which is 832 in our case) [10, 59]. We test the following fractions of the number of input channels: 1/256, 1/128, 1/64, 1/32, 1/16, 1/8, 1/4 and 1/2. Figure 7 shows the result. Even with a low dimensionality in the embedding space, we obtain high activity recognition scores. Note, that even with $K = 3$ for the embedding space with 77.4% accuracy, we outperform the individual video (76.3%) and glove (68.1%) modalities.

**Effect of hand visibility.** Every activity category is associated with a number of hands involved in the activity. Due

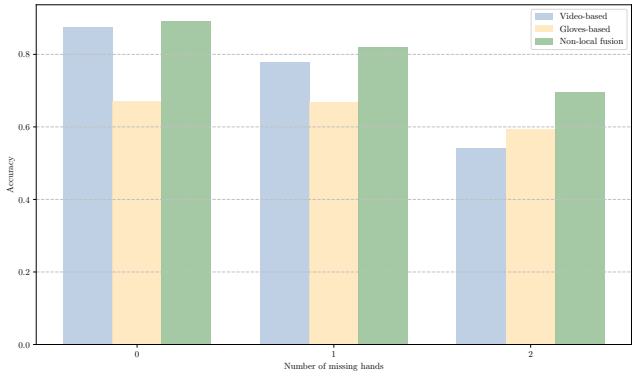


**Figure 8: The effect of hand visibility.** The occluded hands affect the performance of models differently: the video-based model is strongly effected by the visibility of hands, while the gloves-based is logically not effected by the hands visibility. Our non-local fusion can more robustly deal with all scenarios.

to occlusions and varying view points, the number of hands visible in every video clip may be smaller than expected. This might result in miss-classification of samples when using videos only. Here, we investigate how hand visibility affects the behaviour of the models. We do so by estimating the number of missing hands for every sample, which is calculated with help of a RetinaNet [37] trained to detect gloves. For each video frame, between 0 and 2 hands are detected and we assign the most dominant number of hands over the frames to the video. Then we compute the difference between the estimated number of hands and the expected number of hands for every video.

Figure 8 highlights that the biggest improvement in the non-local fusion is achieved for the videos where both hands are occluded, i.e. for two missing hands. Such occlusions can happen due to the nature of the activity or due to the limits of the camera scope. Visual absence of hands effects the video-based model the most. Our non-local fusion improves the most over the video-based model when more hands are missing, highlighting the effect of heterogeneous fusion with smart gloves. The fusion aims to get the best out of both worlds.

### 5.3 Qualitative analysis

Lastly, we perform two qualitative analyses to show what kind of attention is learned across the two modalities, as well as highlighting success and failure cases of the fusion.

**Visualizing cross-modal attention.** Non-local fusion extends self-attention mechanism [53] to multiple modalities, where the relations of two inputs exhibit the features of the leading modality. To visualize that we compute activations of the relation function in the embedding space. Figure 9 shows examples of the most dominant relation regions between the video and the smart gloves. We observe that the relevant regions of hands and interactions are highlighted by the non-local fusion and the time segments of low visual information are suppressed.

**Success and failure cases.** We also show success and failure cases of our approach. Figure 10 provides three success cases, showing that our fusion model can deal with samples where one or both unimodal solutions failed to give a correct prediction. For the first two examples of the *using screwdriver* activity, the fusion model correctly recognizes the category despite the fact that glove information (first example) or visual information (second example) can be insufficient on their own for a correct prediction. The model can even successfully obtain correct predictions when both individual modalities give an incorrect prediction (third example).

The non-local fusion does not always get the best of both worlds. Figure 11 shows two failure cases for *Using screwdriver*. In both examples one of the two modalities obtain the correct activity on their own. When fused however, the combined model makes incorrect predictions, as the fusion is persuaded by the incorrect modality. Overall, we conclude that our non-local fusion can overcome mistakes induced by individual modalities, but does not always predict the desired activity.

## 6 CONCLUSION

In this work, we present a generalized method of fusing two heterogeneous modalities in a non-local fashion - a step forward from a unimodal non-local operation. We demonstrate how our non-local fusion approach can accommodate representations of various sizes and shapes. The effectiveness of non-local fusion is demonstrated on a new dataset, dubbed GloVid, which records both videos and readings from smart gloves. Experimental evaluation shows that our heterogeneous non-local fusion can capture the complementary nature of the vastly different modalities, without compromising their shapes and topologies. We see non-local fusion as a progression towards effective multi-input networks that operate on various information domains.

## ACKNOWLEDGMENTS

This work was funded by Business Finland (grant No. 1660/ 31/ 2018) and the European Unions Horizon 2020 Research and Innovation Programme (grant No. 777222). Special thanks to Clayton Frederick Souza Leite and Xiuyang Li for helping in composing the GloVid dataset.

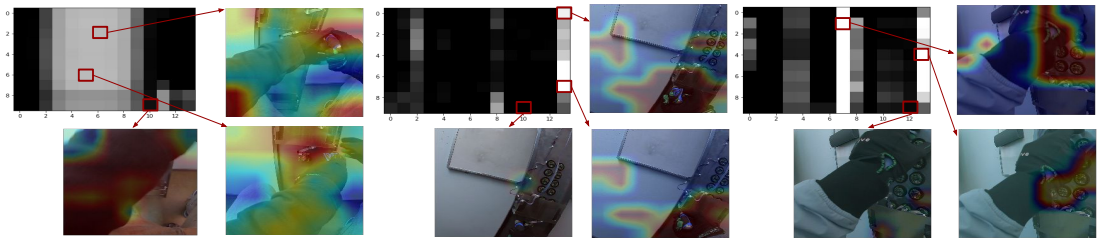

**Figure 9: Examples of activations in the non-local fusion block.** We show the activation matrix relative to video (Y-axis: 10 time points) and smart gloves (X-axis: 14 time points) summed up across the 2D space dimensions of the video. For each two time points we can access the visual information relevancy of the corresponding frame by fixing the time in the output of the function $f$ in Eq. 2 and mapping the activations to the frame. We expand some of the relation points into the heatmaps to show the highlighted information. (Note that absolute values in the time-to-time matrices do not represent the full interaction behaviour due to the summation operation.)


**Figure 10: Success examples of the fusion visualized for the *Using screwdriver* samples.** (top row) Hand-screwdriver interaction is visible, but the glove-based network doesn't recognize the *grabbing* action. (middle row) Interaction is missing from the frame, but hands dynamics allows to correctly recognize the class. (bottom row) Interaction is visible in the first samples, but missing in the remaining ones, fusion model successfully recognizes distributed features of the class.

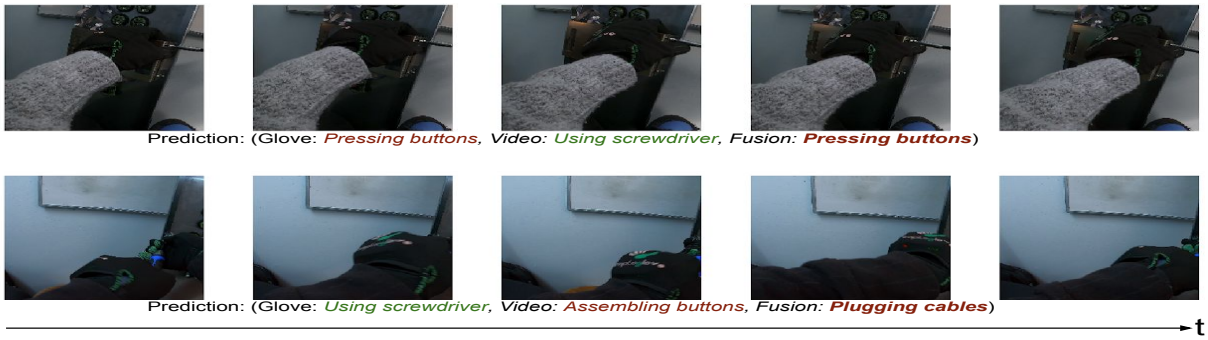


**Figure 11: Failure examples of the fusion visualized for the *Using screwdriver* samples.**(top row) The hand-screwdriver interaction is recognized by the video-based network, but the fusion model is persuaded by the glove recordings. (bottom row) Fusion model fails to use the hands dynamics.

# REFERENCES


[1] Fakhreddine Ababsa, Hicham Hadj-Abdelkader, and Marouane Boui. 2019. 3D Human Tracking with Catadioptric Omnidirectional Camera. In *ICMR*.

[2] A. Akpa, Masashi Fujiwara, Hirohiko Suwa, Yutaka Arakawa, and Keiichi Yasumoto. 2019. A Smart Glove to Track Fitness Exercises by Reading Hand Palm. *Journal of Sensors* (2019).

[3] Xavier Alameda-Pineda, Soraya Arias, Yutong Ban, Guillaume Delorme, Laurent Girin, Radu Horaud, Xiaofei Li, Bastien Mourgue, and Guillaume Sarrazin. 2019. Audio-Visual Variational Fusion for Multi-Person Tracking with Robots. In *ACM MM*.

[4] Serkan Balli, Ensar Arif Saba, and Musa Peker. 2019. Human activity recognition from smart watch sensor data using a hybrid of principal component analysis and random forest algorithm. *Measurement and Control* (2019).

[5] T. Bates, K. Ramirez-Amaro, T. Inamura, and G. Cheng. 2017. On-line simultaneous learning and recognition of everyday activities from virtual reality performances. In *IROS*.

[6] Captoglove. [n.d.]. Captoglove smart gloves' website; https://www.captoglove.com/. https://www.captoglove.com/

[7] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *CVPR* (2017).

[8] Alexandros Andre Chaaraoui, José Ramón Padilla-López, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2014. Evolutionary joint selection to improve human action recognition with RGB-D devices. *Expert systems with applications* 41, 3 (2014), 786–794.

[9] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu. 2012. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics* 42, 6 (Nov 2012), 790–808.

[10] Lu Chi, Guiyu Tian, Yadong Mu, and Qi Tian. 2019. Two-Stream Video Classification with Cross-Modality Attention. In *ICCVw*.

[11] Heeryon Cho and Sang Min Yoon. 2018. Divide and Conquer-Based 1D CNN Human Activity Recognition Using Test Data Sharpening. *MDPI Sensors* (2018).

[12] Pietro Cottone, Gabriele Maida, and Marco Morana. 2014. User activity recognition via kinect in an ambient intelligence scenario. *IERI Procedia* (2014).

[13] Juan Carlos Davila, Ana-Maria Cretu, and Marek Zaremba. 2017. Wearable Sensor Data Classification for Human Activity Recognition Based on an Iterative Learning Framework. *Sensors* (2017).

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

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. 2005. Behavior Recognition via Sparse Spatio-Temporal Features. In *ICCCN*. IEEE Computer Society, USA.

[16] L. Fan, Z. Wang, and H. Wang. 2013. Human Activity Recognition Model Based on Decision Tree. In *CBD*.

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *ICCV*.

[18] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. 2016. Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*.

[19] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. 2017. Spatiotemporal multiplier networks for video action recognition. In *CVPR*. 4768–4777.

[20] C. Feichtenhofer, A. Pinz, and A. Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *CVPR*.

[21] Z. Feng, L. Mo, and M. Li. 2015. A Random Forest-based ensemble method for activity recognition. In *IEEE EMBC*.

[22] Cornelia Fermüller, Fang Wang, Yezhou Yang, Konstantinos Zampogiannis, Yi Zhang, Francisco Barranco, and Michael Pfeiffer. 2018. Prediction of manipulation actions. *IJCV* 126, 2-4 (2018), 358–374.

[23] Korbinian Frank, Mara Jos Vera, Patrick Robertson, and Tom Pfeifer. 2010. Bayesian Recognition of Motion Related Activities with Inertial Sensors. In *UbiComp'10*.

[24] Yuqian Fu, Chengrong Wang, Yanwei Fu, Yu-Xiong Wang, Cong Bai, Xiangyang Xue, and Yu-Gang Jiang. 2019. Embodied One-Shot Video Recognition: Learning from Actions of a Virtual Embodied Agent. In *ACM MM*.

[25] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Multimodal Multimedia Retrieval with vitrivr. In *ICMR*.

[26] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *CoRR* (2017).

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[28] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding.. In *CVPR*.

[29] Francisco Javier Ordez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16 (01 2016), 115.

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*.

[31] Panagiotis Kasnesis, Charalampos Z. Patrikakis, and Iakovos S. Venieris. 2018. PerceptionNet: A Deep Convolutional Neural Network for Late Sensor Fusion. *CoRR* (2018). arXiv:1811.00170

[32] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* (2017).

[33] Heysem Kaya, Furkan Grpnar, and Albert Ali Salah. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing* (2017).

[34] Alexander Klser, Marcin Marszalek, and Cordelia Schmid. 2008. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *BMVC*.

[35] H. Koskimaki, V. Huikari, P. Siirtola, P. Laurinen, and J. Roning. 2009. Activity recognition using a wrist-worn inertial measurement unit: A case study for industrial assembly lines. In *MED*.

[36] Laptev and Lindeberg. 2003. Space-time interest points. In *ICCV*. 432–439 vol.1.

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.

[38] David Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (11 2004), 91–.

[39] Julien Maitre, Clment Rendu, Kvin Bouchard, Bruno Bouchard, and Sebastien Gaboury. 2019. Basic Daily Activity Recognition with a Data Glove. In *Procedia Computer Science*. 108–115.

[40] K. G. Manosha Chathuramali and R. Rodrigo. 2012. Faster human activity recognition with SVM. In *ICTer*. 197–203.

[41] Yang Mi, Kang Zheng, and Song Wang. 2018. Recognizing Actions in Wearable-Camera Videos by Training Classifiers on Fixed-Camera Videos. In *ICMR*.

[42] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. 2018. Moments in Time Dataset: one million videos for event understanding. *IEEE TPAMI* (2018).

[43] Tomokazu Murakami. 2018. Industrial Applications of Image Recognition and Retrieval Technologies for Public Safety and IT Services. In *ACM ICMR*. 4.

[44] Francisco Javier Ordez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016).

[45] Om Prasad Patri, Anand V. Panangadan, Vikrambhai S. Sorathia, and Viktor K. Prasanna. 2016. Sensors to Events: Semantic Modeling and Recognition of Events from Data Streams. *Int. J. Semantic Computing* 10 (2016), 461–502.

[46] Daniel Rotman, Dror Porat, Gal Ashour, and Udi Barzelay. 2018. Optimally Grouped Deep Features Using Normalized Cost for Video Scene Detection. In *ACM ICMR*.

[47] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. 2019. On the Integration of Optical Flow and Action Recognition. In *Pattern Recognition*, Thomas Brox, Andrés Bruhn, and Mario Fritz (Eds.). Springer International Publishing, Cham, 281–297.

[48] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*.

[49] Xing Su, Hanghang Tong, and Ping Ji. 2014. Activity Recognition with Smartphone Sensors. *Tsinghua Science and Technology* 19 (06 2014), 235–249.

[50] Veralia Gabriela Snchez and Nils-Olav Skeie. 2018. Decision Trees for Human Activity Recognition in Smart House Environments. In *SIMS*. Linkping University Electronic Press, Linkpings universitet.

[51] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2017. A Closer Look at Spatiotemporal

Convolutions for Action Recognition. *CoRR* (2017).
[52] Gül Varol, Ivan Laptev, and Cordelia Schmid. 2017. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1510–1517.
[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
[54] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Multi-Level Sensor Fusion with Deep Learning. *CoRR* (2018).
[55] Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal Multimodal Fusion for Video Emotion Classification in the Wild. *CoRR* (2017).
[56] Robert-Andrei Voicu, Ciprian Dobre, Lidia Bajenaru, and Radu-Ioan Ciobanu. 2019. Human Physical Activity Recognition Using Smartphone Sensors. *Sensors* 19, 3 (2019).
[57] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. 2009. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009 - British Machine Vision Conference*, A. Cavallaro, S. Prince, and D. Alexander (Eds.). BMVA Press, London, United Kingdom, 124.1–124.11.
[58] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. 2018. Appearance-and-relation networks for video classification. In *CVPR*.
[59] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2017. Non-local Neural Networks. *CoRR* (2017).
[60] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. 2008. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In *ECCV*, David Forsyth, Philip Torr, and Andrew Zisserman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 650–663.
[61] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. 2007. A Scalable Approach to Activity Recognition based on Object Use. In *ICCV*. 1–8.
[62] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 305–321.
[63] Juan Ye, Graeme Stevenson, and Simon Dobson. 2016. Detecting abnormal events on binary sensors in smart home environments. *Pervasive and Mobile Computing* 33 (06 2016).
[64] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Proc. and Control* 47 (2019), 312–323.
[65] Yu Zhao, Rennong Yang, Guillaume Chevalier, and Maoguo Gong. 2017. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors. *CoRR* (2017).