future internet

MDPI

Article

Multimodal Deep Learning for Group Activity Recognition in Smart Office Environments

George Albert Florea [1] and Radu-Casian Mihailescu [1,2,*]

[1] Department of Computer Science, Malmö University, 20506 Malmö, Sweden; george–albert@hotmail.com
[2] Internet of Things and People Research Center, Malmö University, 20506 Malmö, Sweden
* Correspondence: radu.c.mihailescu@mau.se

Received: 29 June 2020; Accepted: 4 August 2020; Published: 9 August 2020

Abstract: Deep learning (DL) models have emerged in recent years as the state-of-the-art technique across numerous machine learning application domains. In particular, image processing-related tasks have seen a significant improvement in terms of performance due to increased availability of large datasets and extensive growth of computing power. In this paper we investigate the problem of group activity recognition in office environments using a multimodal deep learning approach, by fusing audio and visual data from video. Group activity recognition is a complex classification task, given that it extends beyond identifying the activities of individuals, by focusing on the combinations of activities and the interactions between them. The proposed fusion network was trained based on the audio–visual stream from the AMI Corpus dataset. The procedure consists of two steps. First, we extract a joint audio–visual feature representation for activity recognition, and second, we account for the temporal dependencies in the video in order to complete the classification task. We provide a comprehensive set of experimental results showing that our proposed multimodal deep network architecture outperforms previous approaches, which have been designed for unimodal analysis, on the aforementioned AMI dataset.

Keywords: multimodal learning; deep learning; activity recognition

1. Introduction

Activity recognition is nowadays an active research topic, with ramifications over numerous application domains. To name a couple, activity recognition plays an important role in providing personalized support for healthcare monitoring systems [1] and assisted living for elderly care [2]. Additionally, as energy management is regarded the highest operational expenditure in commercial and residential buildings [3], several activity-based solutions have already been proposed to optimize the use of energy in smart home and building environments, by providing the adequate level of automation through dedicated services (e.g., [4]).

Advances in this area are primarily fostered by the ever-increasing number of sensing devices that are permeating our surrounding environment. At the same time, activity recognition is attracting evermore attention from the field of machine learning which is rapidly expanding, oftentimes as a result of new and specific problem definitions driven by practical applications. In this regard, we put forward in this work a deep learning-based approach, which can model and account for different input modalities by fusing the data in order to perform activity recognition in the context of office environments. Our empirical results were validated based on the AMI Corpus dataset [5], which has been put forward by a European consortium that is developing meeting browsing technology (see illustration in Figures 1 and 2). In particular, the multimodal approach introduced in this paper was designed to integrate the audio–visual cues in order to distinguish among several typically occurring activity classes. The workplace environment is undergoing considerable transformations

aimed towards fostering agile and flexible work, by means of enabling the environment to adapt and contextualize the experience to the needs and preferences of its users. This can range from controlling things such as heating, lighting, and ventilation in order to increase user satisfaction (e.g., [6]), to more complex scenarios, wherein a dynamic set of devices, with their functionalities and services, cooperate temporarily to achieve a user's goal (e.g., [7]). It is therefore paramount to provide mechanisms that can lead to improved performance in activity recognition tasks, which can then, in turn, communicate this information to various computational decision nodes in the environment with the goal of assisting office users or maintenance personal with their daily activities.

(**a**) Corner camera angle (**b**) Overhead camera angle

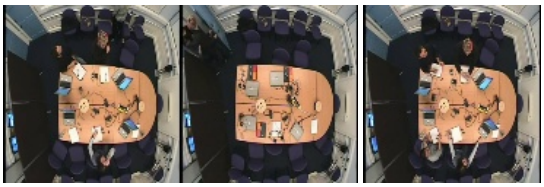
Figure 1. Images showing different camera angles from the AMI Corpus dataset.

The rest of the paper is organized as follows. In Section 2 we introduce the relevant background and related work. Section 3 introduces the proposed model design for activity recognition in smart office environments. Section 4 presents the evaluation results and discusses the performance of our approach against previous work on the AMI dataset. The paper concludes in Section 5.

2. Background

2.1. Activity Recognition in Smart Buildings

The majority of related work in the area of activity recognition for office space and building environments has largely been focused so far on occupant presence detection. Inferring this information alone has been shown to have a great impact in terms of optimizing the energy consumption based on adapting the heating, cooling, and ventilation systems in response to occupancy data. For instance, in [8] the authors deployed a sensor network, consisting of plug meters for measuring appliance energy consumption, light sensors, and PIR sensors that were triggered by human motion. A Gaussian mixture model with 97% accuracy was used for the task of determining whether a user was present in the office or not. Based on this result the authors concluded that 35.5% potential savings could be achieved by simply turning off the appliances and lightning when the office was empty. Similarly, in [4] the authors proposed a mechanism based on detecting activity patterns in smart meeting rooms, coupled with a scheduling algorithm designed to coordinate the heat supply accordingly. Other solutions, instead of resorting to sensor deployments in the environment, proposed capturing data from the user's smartphone device in order to recognize different activities. A notable examples is reported in [9], wherein the authors collected accelerometer data for determining floor localization within a building without any prior knowledge about the premise. Alternatively, in this work we are concerned with processing audio–visual data streams from office environments for group activity recognition. Group activity recognition is a complex classification task, given that it extends beyond identifying the activities of individuals, by focusing on the combinations of activities and interactions between them. Thus, we hypothesized that the audio and visual modalities are a suitable choice for this task.



Presentation Empty/No-activity Meeting

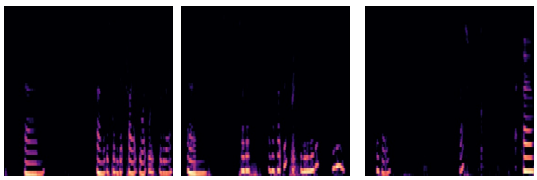
Figure 2. Examples from the AMI Corpus dataset for each of the three classes under consideration.

2.2. Deep Learning for Activity Recognition

The advent of deep learning models has seen neural networks being successfully applied across many classification problems, with implications especially in the area of image processing and analysis. Since its emergence in 2009, it has consistently replaced state-of-the-art approaches to machine learning and largely overcome the problem of hand-crafted feature extraction that dominated the field precedently.

In [10] the authors propose a framework that couples skeleton data extracted from RBG video and a deep bidirectional-LSTM model for activity recognition. The limited information in this instance is the lack of depth information, as opposed to RGB-D video, which includes depth data. However, acquiring depth information is costly and requires special equipment. The authors propose a set of techniques meant to augment the RGB dataset, such as dynamic frame dropout and gradient injection, which eventually was shown to outperform the state-of-the-art RGB-D video stream solutions and can therefore be widely deployed using ordinary cameras. Although it provides several valuable augmentation techniques, this work only dealt with video data.

In [11] the authors explore different approaches in multi-sound classification and propose a stacked classifier based on recent advancements in deep learning. The proposed approach can robustly classify sound categories among mixed acoustic signals, without the need for prior knowledge about the numbers and signatures of sounds in the mixed signals. The performance achieved was around 60% in terms of the F1-score, depending on the number of mixed sounds. The proposed model was trained based on a synthesized sound effect dataset, which is a clear limitation in contrast to real-world data which tends to present rather different acoustic signatures. As can be observed in Figure 3, the spectogram representation of the audio signal can be difficult to interpret, and even situations where a room is empty may exhibit background noise, which could hinder the classification. The deep learning model proposed here is only designed to address audio data streams.



Presentation Empty/No-activity Meeting

Figure 3. Mel audio spectogram corresponding to the activities in Figure 2.

2.3. Multimodal Learning

Our work belongs to the class of multimodal methods, wherein models are built to process and relate information from multiple sensing modalities. There are several multimodal architectures specifically designed for the task of activity recognition. Depending on the structure according to which the unimodal approaches are integrated (such as the ones outlined in Section 2.2), these can be classified into three main categories: *data fusion*, *feature fusion*, and *decision fusion*. In the case of audio–visual classifiers, as it pertains to our work, data fusion is rarely an option due to the fact that the raw image and audio data are incompatible.

In [12] the authors were concerned with integrating heterogeneous and distributed sensors in a dynamic manner, thereby introducing the concept of dynamic intelligent virtual sensors (DIVS) in

order to support the creation of services designed to tackle complex problems based on reasoning about various types of data. Moreover, they presented an application of DIVS in the context of activity monitoring using sensors in a smart building environment. However, the authors focused on *decision-level fusion* techniques, while assuming that local processing units performed the low-level processing (e.g., image recognition), and then those classification decisions were further synthesized to obtain a final decision. The multi-stream neural network architecture introduced in [13] belongs to the same class of *decision-level fusion* models. Spatial and temporal streams are constructed for each modality. The fusion method then aggregates the classification decisions from each stream by determining the corresponding fusion weights for generating the final scores of each class. In addition, relations between classes are used in the classification process.

Feature-level fusion is applied in [14] by combining audio–visual features with multi-kernel learning and multi-kernel boosting. After extracting specific features for each modality, the proposed framework performs adaptive fusion by selecting and weighting the features and kernels jointly. This work tackles the problem of egocentric activity recognition, which consists of analyzing first-person videos, wherein the world is seen from the perspective of the actor whose actions need to be recognized. A similar example of egocentric activity recognition is introduced in [15]. In contrast, the work presented in this paper, though also resorting to *feature-level fusion*, focuses on group activity recognition that involves multiple actors. A common practice concerning *feature-level fusion* for video data is to compute temporal aggregation for each individual modality before features are fused. For instance, temporal segment networks [16] are capable of modeling a long-range temporal structure over an entire video, by dividing the video into several segments and then fusing them based on a segmental consensus function for each modality. Lastly, classifications from all modalities are fused to produce the final classification. Alternatively, as will be evident in Section 3.3, in this work we first performed audio–visual *feature fusion*, which was followed by temporal aggregation, thereby reducing the size of the network.

Most similarly to our work, in [17], the authors investigated deep learning architectures for the task of activity recognition based on the AMI dataset. Three activity classes were used: *presentation*, *meeting*, and *empty*, as depicted in Figure 2. In addition, the paper provides insights with regard to the benefits of applying transfer learning in solving activity recognition problems. Further, the authors demonstrate that high performance can be preserved even when the model is deployed to a real-word laboratory environment. In this paper we build upon the previous work in [17]. However, we are introducing here a multimodal approach based on residual learning that allows one to fuse audio and visual data from video, which goes beyond the unimodal method in [17].

3. Deep Multimodal Architecture

In the following section we provide the details of the problem domain and our proposed approach for activity recognition in office environments and present the structure of the multimodal deep network.

The area of activity recognition in office environments is rather complex as a computer vision classification problem. In contrast to the problem of object recognition, wherein the goal is to identify distinct objects in images, activity recognition from video data requires scene understanding over a certain time horizon, wherein subjects are moving and interacting. For instance, in order to distinguish between a *presentation* and a *meeting*, it is not enough to identify the activities of individual users; one must also identify the combinations of activities and interactions between them. Moreover, the presence and detection of certain objects in the scene can inform and help with discerning between different activities. Using a whiteboard, for example, would provided different clues regarding the ongoing activity, as opposed to working on a laptop.

The proposed procedure for activity recognition essentially consists of two steps. First, we need to extract a joint audio–visual feature representation suitable for activity recognition. To this end, our deep network is composed of two streams. On the one hand, a visual model is used for processing the visual cues, and on the other hand, an audio model is responsible for processing the audio signal.

Second, a fusion network combines the visual and audio extracted features. To this end, we use an LSTM network [18] to model the spatio-temporal evolution of the feature representation over time in order to perform the classification on video segments. For understanding the behavior and interactions of objects and persons in the scene, it is key to model the spatio-temporal progression of the audio–visual data.

3.1. Visual Network

Due to computational considerations, the input to the network consisted of a video feed of size 144×144, downsampled from the original higher resolution of 720×576 provided by the AMI dataset. The video signals were stored on the disk using DivX AVI codec 5.2.1. The encoding bit rate was 23,000 Kbps, with a maximum interval of 25 frames between two consecutive MPEG keyframes, in order to reduce redundant information. In contrast to the previous work in [17], the basis of our visual model is represented by a deep residual network (ResNet [19]). Residual learning has emerged as a way to combat the problem of vanishing or exploding gradients in deep networks with more than 30 hidden layers, by stacking building blocks that introduce so-called residual connections of the form:

$$y = \mathcal{F}(x, \{W\}) + h(x) \tag{1}$$

where x and y represent the input and output; $h(x)$ stands for the identity mapping or a linear projection to match the dimension of $\mathcal{F}$, which is the residual function to be learned. In comparison with other popular network architectures, such as All-CNN [20], DenseNet [21], FractalNet [22], and Highway Network [23], ResNet typically achieves state-of-the-art or similar results, while using fewer parameters, on typical image processing and classification tasks [21]. In particular, we mainly use ResNet-50, which consists of 5 main bottlenecks, each including a convolution and a residual block. Each convolution block has 3 convolution layers. The same is the case for each residual block. Overall, the ResNet-50 has over 23 million trainable parameters.

3.2. Audio Network

The audio signal available in the AMI dataset was downsampled from 48 kHz to 16 kHz (which corresponds to a 96,000-dimensional vector), and made available in a standardized format as WAV files for each meeting. In addition, the audio signal was smoothed out with a low-pass filter and an automatic energy threshold was applied to distinguish between speech or silence. Now, in order to convert the one dimensional audio signal into a viable input for our model, we extracted Mel spectograms from the audio files (see Figure 3). Specifically, the FFT was configured with a time span of 25 ms and a window hop of 10 ms, and was computed over each frame of the original time-domain signal. The result was mapped to 64 Mel bins and the frequencies that were cut off were in the [175, 7500] Hz range. Then, in order to stabilize the spectogram, we took the logarithm. The resulting log Mel spectogram further used for feature extraction had a 144×144 size. Similarly to the visual model presented in the previous section, we applied deep residual learning for processing the spectograms.

3.3. Temporal Audio–Visual Fusion Model

The segmented video sequences used in training were decoupled into synchronized audio and visual data $X = \{x_{a_i}, x_{v_i}, y_i\}_{i=1,k}$, where x_{a_i} and x_{v_i} represent the audio and video examples respectively, and the corresponding activity class is denoted by y_i. Let Γ denote the ResNet-based feature extractor that maps the input to a D-dimensional feature vector, such that $f = \Gamma(x; \theta)$, where the parameters of this mapping are given by the vector θ. It follows that the audio network can then be denoted by $\Gamma^a(x_a; \theta_a) = f^a$, and the visual network as $\Gamma^v(x_v; \theta_v) = f^v$. During the first phase of the training procedure we resort to separately training the audio (Γ^a) and visual (Γ^v) models in order to learn more discriminative audio and visual feature representations, by minimizing the label

prediction loss on each network individually. Both the audio and visual networks extract, respectively, $D = 1024$-dimensional feature vectors.

In the second phase, we discard the softmax layer in Γ^a and Γ^v and introduce an LSTM-based fusion network. The higher-layer representations that are extracted from both the audio and visual streams provide structural information that is further fused by the LSTM to model the spatio-temporal evolution over time. At the same time, this step ensures that contextual information can be captured from the data. Specifically, we use two uni-directional LSTM layers with 128 cells each to sequentially learn based on the joint audio–visual feature representation. While this step is being performed, i.e., while the LSTM layers are being trained, the audio and visual networks keep their parameters fixed. This is achieved by optimizing with backprop the following minimization problem:

$$\min_{\theta_F} \sum_{i=1}^{K} L(\Gamma^F([\Gamma^v(x_{v_i};\theta_v), \Gamma^a(x_{a_i});\theta_a];\theta_F)), y_i) \tag{2}$$

Here, Γ^F and θ^F represent the fusion network and the corresponding network parameters, while L denotes the softmax log loss function, which is calculated by:

$$L(\Gamma, y) = -\sum_{j=1}^{m} y_j log(y_j^{\Gamma}) \tag{3}$$

where m is the total number of classes to be recognized, the ground truth class label for the j-th example is y_j, and y_j^{Γ} denotes the j-th output of the softmax layer of the Γ network. In addition, for regularization of the network, given the large number of parameters, we adopted the droput method with $p = 0.5$ and batch normalization. During training we used small mini-batches of size 32 due to hardware limitations, as the system was implemented on the Google Colab platform.

4. Empirical Results

In this section we describe the experimental setup for our analysis and report results based on our deep learning approach for the task of activity recognition in smart office environments, which were equipped with some form of audio–visual monitoring system. As previously emphasized, for this study we employed the AMI dataset, which we briefly describe in the following section. For all experiments we considered a 80/20 ratio between training and validation data and used the ResNet-50 architecture as the base model, when not specified otherwise. The specifics of the audio–visual data are described in Sections 3.1 and 3.2 as inputs for the respective networks. The AMI Corpus NITE XML Toolkit (http://groups.inf.ed.ac.uk/nxt/) was used to extract the labels for each video. In order to evaluate the efficiency of our approach, besides the loss function introduced in Equation (3), for an easier interpretability we also report the following metrics:

- *Classification accuracy*, which is the ratio of correct predictions to the total number of input samples:

$$Acc = \frac{\sum_{x_i \in X_{text}} 1\{f(x_i) = y_i\}}{|X_{text}|} \tag{4}$$

- *F1-score*, which is a measure of both the precision of the classifier and its robustness, defined as the harmonic mean between precision and recall:

$$F1 = 2 * (Recall * Precision)/(Recall + Precision) \tag{5}$$

In the following, we describe a comprehensive empirical study, whereby we first investigated the classification task in the context of each separate modality, exploring our approach against several DL base models and different model complexities. Second, we analyzed the results on multimodal data and evaluated the advantages of transfer learning in this context.

4.1. Dataset

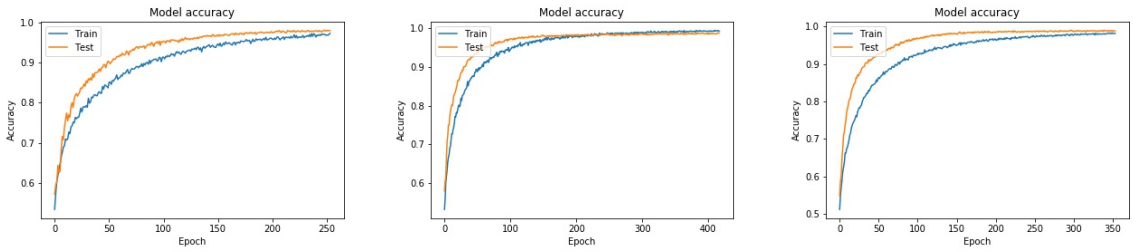
The AMI Corpus is a European-funded project that aims to improve the effectiveness of meetings by making it possible to analyze recordings of meetings in order to suggest technologies that can improve the overall experience. The AMI dataset consists of 100 h of meeting recordings, which are organized into 10–60 min long synchronized video and audio sequences. These sequences took place in an office environment where the participants were not always the same people; however, there were always four people present, each given one of four different roles. The meetings are annotated with different labels, which for the purpose of our experiments, were categorized to correspond to one of following three classes: *presentation*, *meeting*, or *empty* (see Figure 2). The videos used in this work were those wherein the camera angle captured the entire room, namely, the overhead view (top view from above) and corner view (camera positioned at the top or a corner), as depicted in Figure 1, which correspond to the Edinburgh and Idiap scenario meetings. Audio data were obtained from far-field microphones and room-view video cameras. The meetings were recorded in English using three different rooms with different acoustic properties, and included mostly non-native speakers. Special hardware was used to provide synchronization for the audio and visual data. The recordings used a range of signals synchronized to a common timeline. The AMI dataset was also used for investigating applications such as speaker role labeling (e.g., [24]) and meeting summarization (e.g., [25]).

4.2. Unimodal Results Analysis

In order to gain a better understanding of the added value obtained via the deep residual learning approach the first set of experiments focused on results that can be attained by treating each modality individually.

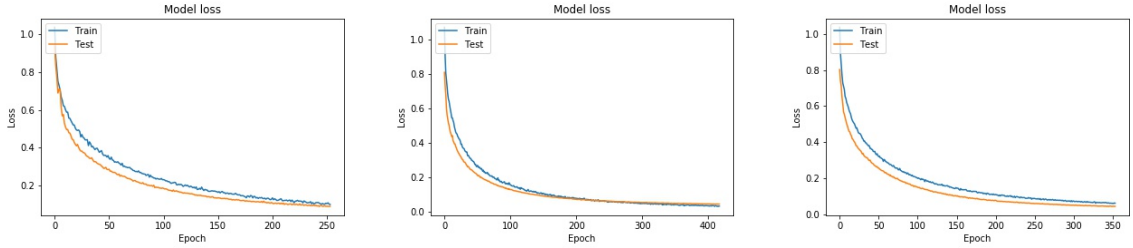
4.2.1. Visual Modality

As a starting point we set out to investigate the effects of different camera angles relatively on the activities occurring. In Figures 4 and 5 we plotted the model's accuracy and loss respectively over the course of training. It is interesting to note that regardless of the camera's positioning and field of view, both the overhead and the corner camera achieved a high accuracy; however, the learning curve slope for the overhead camera is slightly steeper than in the case of the corner camera. That is, while the overhead camera achieved 99% accuracy after about 200 epochs, the corner one got to 98% in roughly 250 epochs. In addition, when activity recognition was performed based on video streams from both camera angles, the resulting accuracy was also 99%. Thus, it appears that the overhead camera provided the best field of view for addressing activity recognition in an office setup, while the corner angle did not provide any performance increase. Importantly, in contrast to the previous work [17] on the AMI dataset, our residual learning-based approach outperformed by a margin of 5%.



(**a**) (**b**) (**c**)

Figure 4. Model accuracy based on the visual data from (**a**) the corner camera, (**b**) the overhead camera, and (**c**) the corner and overhead camera angles combined.



(**a**) (**b**) (**c**)

Figure 5. Model loss based on the visual data from (**a**) the corner camera, (**b**) the overhead camera, and (**c**) the corner and overhead camera angles combined.

Given that the two classes *presentation* and *meeting* clearly display a level of similarity much higher compared to the *empty* class, which may impact the overall results, we also conducted a separate experiment wherein the classification task was strictly limited to distinguishing between *presentation* and *meeting*. The accuracy remained the same; Figure 6 presents the confusion matrix, which indicates that the model is more likely to misclassify a meeting as presentation than vice-versa.

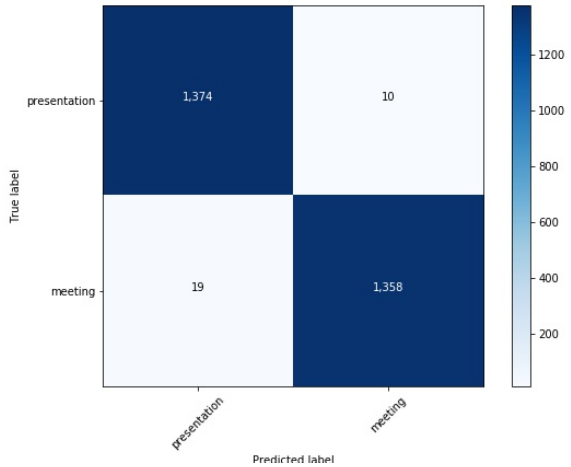
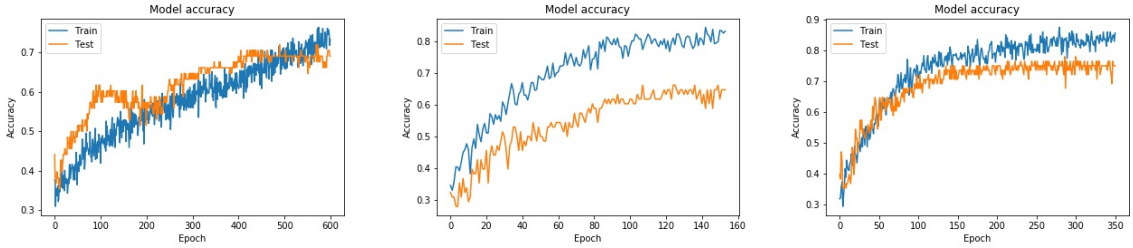


Figure 6. Confusion matrix for the corner and overhead camera angles combined.

We set out to investigate the performance for the activity classification task based solely on the audio feed. Namely, in Figure 7 we compare several deep learning architectures that showed promising results in [17] on the AMI dataset, against the residual learning approach. Noticeably, as well as in the case of visual data, residual learning outperformed both the VGG-16 [26] and the Inception V3 [27] models, by extracting features that made it easier for the model to classify the audio data correctly. Moreover, the differences in the F1-score are considerably larger, as can be seen in Table 1. At the same time, it is important to point out that audio-based activity recognition did not perform anywhere close to the accuracy attained when using visual data, which is consistent with similar findings reported for activity recognition tasks [16,28]. Additionally, we mentioned that the results reported here have been obtained based on a transfer learning approach, which is further detailed in Section 4.2.3.



(**a**) (**b**) (**c**)

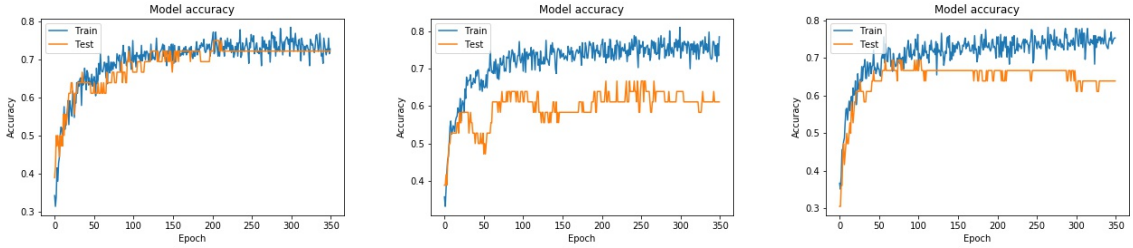
Figure 7. Model accuracy for each of the three base models on the audio data set: (**a**) VGG-16, (**b**) Inception V3, and (**c**) ResNet-50.

Table 1. F1-score results of activity recognition for different modalities.

Visual Spatial Features		**Audio Spatial Features**		**Audio Spatial Features**		**Spatio-Temporal Features**	
Corner	98%	VGG-16	50%	ResNet-50	78%	Audio features	59%
Overhead	99%	Inception V3	66%	ResNet-101	69%	Visual features	100%
Combined	99%	ResNet-50	78%	ResNet-152	66%	Fused features	100%

4.2.2. Deeper Networks

Given the results obtained on the audio dataset, we further investigated the extent to which performance can benefit from deeper architectures. To this end, in Figure 8 we present the accuracy achieved when the base feature extractor model had 50, 101, or 152 hidden layers respectively. It is interesting to remark that although the expected result should have been to obtain a higher accuracy, as the complexity of the model was increased, that was not actually the case. In fact, it is clear from Figure 8 that deeper networks tend to overfit the data, leading to lower accuracy on the validation set. However, one possible explanation is that this situation occurred due to the size of the dataset, and some performance gain could still be achieved, provided we increase the size of the training set.



(**a**) (**b**) (**c**)

Figure 8. Model accuracy for different base models: (**a**) ResNet-50, (**b**) ResNet-101, and (**c**) ResNet-152.

4.2.3. Transfer Learning

Finally, for the audio dataset we provide a comparative analysis between the following two instances. On the one hand, we evaluate the performance of the model when the network weights were randomly initialized, and on the other hand, we compare it against the results in Section 4.2.4, where we have adopted a transfer learning approach by means of using weights that were pre-trained on the ImageNet dataset [29], which is a typical benchmark for solving image recognition problems. This transfer learning approach enables us to use a large-scale dataset, such as ImageNet, as a feature extractor for low-level structures that occur in most images. After initializing the network with the pre-trained weights, the model was further fine-tuned for the activity recognition classification task under consideration, instead of learning it from scratch. As depicted in Figure 9, we can see a significant drop in accuracy for the ResNet-50 base model, which only managed to reach 56% accuracy, in contrast to the 78% that was attained when transfer learning was applied (see Figure 7c). Hence, this

is a strong indication of the fact that transfer learning can boost performance for the task of activity recognition, even on the audio dataset.

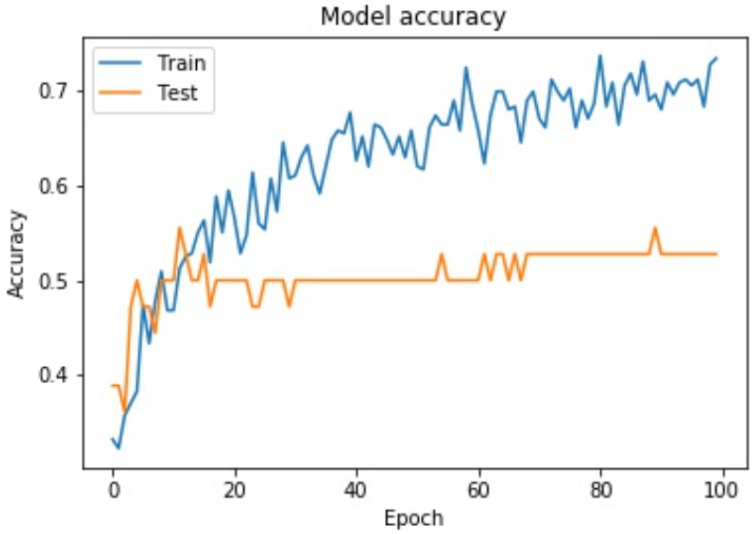


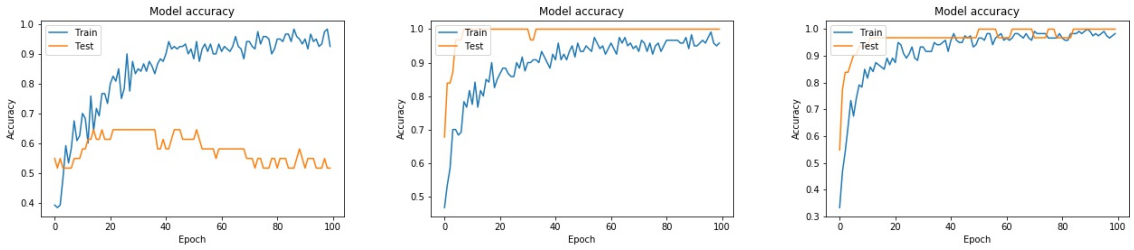
Figure 9. Model accuracy with random weight initialization.

4.2.4. Audio Modality

4.3. Temporal Multimodal Results Analysis

Now that we have explored the limitations of the unimodal approaches based on spatial features only, we focus on the proposed multimodal architecture, which also incorporates temporal features. As already explained in Section 3.3, we appended the network architecture by stacking two LSTM layers that also have the role of fusing the spatio-temporal features extracted from the two respective networks.

For completeness of results we also provide in Figure 10 the performances of the audio and visual models separately, when temporal features were considered. In terms of the audio model, the performance appears to remain relatively low even as temporal features are being incorporated. However, it is interesting to observe that both the visual model and the multimodal one, which fuses the audio and visual streams, managed to achieve 100% accuracy on the AMI dataset. Moreover, this was attained in less than 20 epochs of training, as opposed to the results in Section 4.2, which converged after 200 epochs.



(**a**) (**b**) (**c**)

Figure 10. Model accuracy including spatio-temporal features (**a**) only audio, (**b**) only video, and (**c**) both audio and video combined.

4.4. Discussion

As opposed to previous work on the AMI dataset, in this paper we were particularly interested to analyze the extent to which audio data could be utilized for activity classification. Although audio

monitoring may raise even greater privacy concerns than visual data, it has the advantage of being less dependent on the environment in the same way that video is. That is, the generated audio spectograms are not affected by the different camera angles, nor by people behaving temporarily in a way that does not fit with the annotated label, or other varying conditions such as camera occlusions and lighting. Both camera angles share the same audio feed, so in this regard, using audio seems to be more beneficial compared to video. Nevertheless, audio also has its drawbacks, such as that the same audio pattern could potentially be recorded for entirely different visual behaviors, which could in fact reflect different activities. With those considerations in mind, we have introduced here a multimodal approach that can benefit from the complementary information encoded by both of these two modalities. As a result, the model was shown to achieve 100% accuracy for the specific classification task. Moreover, we have shown that by adopting residual learning we can further improve on the unimodal results in [17] by a considerable margin, and the particular choice for the number of hidden layers in the network impacts performance significantly.

Finally, using the ImageNet weights instead of randomly initializing the base models provides an advantage regarding the amount of time needed to train the model. The base models are quite large with a lot of parameters, and randomly initializing the weights of the base model would take a lot of resources in order to optimize the weights for extracting the features. As can be observed from the results, by applying a transfer learning method we have largely improved performance. This becomes especially important in situations wherein the training set is of limited size.

5. Conclusions and Future Work

In this paper we propose a multimodal approach that operates on audio–visual data streams to perform activity recognition related to office environments. To train the model, we employ a two-step procedure. First, we apply a transfer learning method to initialize and fine-tune the audio and visual deep residual networks. Second, the LSTM-based fusion network is trained to extract a joint audio–visual feature representation, for considering contextual information in the data. The model was evaluated based on the AMI dataset, which consists of 100 h of synchronized video and audio sequences from meeting recordings. Empirical results show that our model achieves significantly better performance in comparison to previous models using the AMI dataset. As future work, we would like to evaluate our approach on a larger number of activities, with a broader range of behavioral and movement patterns, over a longer period of time. This could be achieved by either collecting more data in order to expand the AMI dataset, or by repurposing already existing video databases (such as relevant video sequences from YouTube) for our objective.

Author Contributions: Conceptualization, G.A.F. and R.-C.M.; Methodology, G.A.F. and R.-C.M.; Software, G.A.F. and R.-C.M.; Validation, G.A.F. and R.-C.M.; Formal analysis, R.-C.M.; Investigation, G.A.F.; Data curation, G.A.F.; Writing–original draft preparation, R.-C.M.; Writing–review and editing, G.A.F. and R.-C.M.; Visualization, G.A.F.; Supervision, R.-C.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Stiftelsen för Kunskaps- och Kompetensutveckling grant number 20140035.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Islam, S.M.R.; Kwak, D.; Kabir, M.H.; Hossain, M.; Kwak, K. The Internet of Things for Health Care: A Comprehensive Survey. *IEEE Access* **2015**, *3*, 678–708. [CrossRef]
2. Chernbumroong, S.; Cang, S.; Atkins, A.; Yu, H. Elderly activities recognition and classification for applications in assisted living. *Expert Syst. Appl.* **2013**, *40*, 1662–1674. [CrossRef]
3. Minoli, D.; Sohraby, K.; Occhiogrosso, B. IoT Considerations, Requirements, and Architectures for Smart Buildings—Energy Optimization and Next-Generation Building Management Systems. *IEEE Internet Things J.* **2017**, *4*, 269–283. [CrossRef]

4. Lim, B.; Briel, M.; Thiébaux, S.; Backhaus, S.; Bent, R. HVAC-Aware Occupancy Scheduling. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, Austin, TX, USA, 25–30 January 2015; AAAI Press: Palo Alto, CA, USA, 2015; pp. 679–686.
5. Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; et al The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*; Renals, S., Bengio, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 28–39.
6. Truong, N.C.; Baarslag, T.; Ramchurn, G.; Tran-Thanh, L. Interactive scheduling of appliance usage in the home. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-160 (15/07/16), New York, NY, USA, 9–11 July 2016.
7. Yang, Y.; Hao, J.; Zheng, Y.; Yu, C. Large-Scale Home Energy Management Using Entropy-Based Collective Multiagent Deep Reinforcement Learning Framework. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 630–636.
8. Ahmadi-Karvigh, S.; Ghahramani, A.; Becerik-Gerber, B.; Soibelman, L. Real-time activity recognition for energy efficiency in buildings. *Appl. Energy* **2018**, *211*, 146–160. [CrossRef]
9. Ye, H.; Gu, T.; Zhu, X.; Xu, J.; Tao, X.; Lu, J.; Jin, N. FTrack: Infrastructure-free floor localization via mobile phone sensing. In Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications, Lugano, Switzerland, 19–23 March 2012; pp. 2–10. [CrossRef]
10. Sarker, K.; Masoud, M.; Belkasim, S.; Ji, S. Towards Robust Human Activity Recognition from RGB Video Stream with Limited Labeled Data. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018. [CrossRef]
11. Haubrick, P.; Ye, J. Robust Audio Sensing with Multi-Sound Classification. In Proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications, Kyoto, Japan, 11–15 March 2019; pp. 1–7.
12. Mihailescu, R.C.; Persson, J.; Davidsson, P.; Eklund, U. Towards Collaborative Sensing using Dynamic Intelligent Virtual Sensors. In *Intelligent Distributed Computing*; Badica, C., El Fallah Seghrouchni, A., Beynier, A., Camacho, D., Herpson, C., Hindriks, K., Novais, P., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 217–226.
13. Wu, Z.; Jiang, Y.G.; Wang, X.; Ye, H.; Xue, X. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification. In Proceedings of the 24th ACM International Conference on Multimedia, MM '16, Amsterdam, The Netherlands, 15–19 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 791–800.
14. Arabacı, M.A.; Özkan, F.; Surer, E.; Jančovič, P.; Temizel, A. Multi-modal egocentric activity recognition using multi-kernel learning. *Multimed. Tools Appl.* **2020**. [CrossRef]
15. Kazakos, E.; Nagrani, A.; Zisserman, A.; Damen, D. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 5491–5500.
16. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 20–36.
17. Casserfelt, K.; Mihailescu, R. An investigation of transfer learning for deep architectures in group activity recognition. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019, Kyoto, Japan, 11–15 March 2019; pp. 58–64. [CrossRef]
18. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
20. Springenberg, J.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net. *arXiv* **2014**, arXiv:1412.6806.
21. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.

22. Larsson, G.; Maire, M.; Shakhnarovich, G. FractalNet: Ultra-Deep Neural Networks without Residuals. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
23. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Training Very Deep Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2*; MIT Press: Cambridge, MA, USA, 2015; pp. 2377–2385.
24. Sapru, A.; Valente, F. Automatic speaker role labeling in AMI meetings: Recognition of formal and social roles. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 5057–5060.
25. Zhao, Z.; Pan, H.; Fan, C.; Liu, Y.; Li, L.; Yang, M.; Cai, D. Abstractive Meeting Summarization via Hierarchical Adaptive Segmental Network Learning. In Proceedings of the World Wide Web Conference, WWW '19, San Francisco, CA USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 3455–3461. [CrossRef]
26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
27. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
28. Lopes, J.; Singh, S. Audio and Video Feature Fusion for Activity Recognition in Unconstrained Videos. In *Intelligent Data Engineering and Automated Learning—IDEAL 2006*; Corchado, E., Yin, H., Botti, V., Fyfe, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 823–831.
29. 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. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]