International Journal of Neural Systems, Vol. 23, No. 5 (2013) 1350021 (15 pages)

DOI: 10.1142/S0129065713500214

World Scientific
www.worldscientific.com

MULTIMODAL INTEGRATION OF MICRO-DOPPLER SONAR AND AUDITORY SIGNALS FOR BEHAVIOR CLASSIFICATION WITH CONVOLUTIONAL NETWORKS

SALVADOR DURA-BERNAL
Department of Physiology and Pharmacology
SUNY Downstate, 450 Clarkson Avenue
Brooklyn, NY 11203, USA
salvadordura@gmail.com

GUILLAUME GARREAU* and JULIUS GEORGIOU†
Holistic Electronics Research Lab, University of Cyprus
75 Kallipoleos Street
Nicosia, 1678 Nicosia, Cyprus
**garreau.guillaume@ucy.ac.cy*
†*julio@ucy.ac.cy*

ANDREAS G. ANDREOU
Electrical and Computer Engineering
Johns Hopkins University, Barton Hall 400
3400 N. Charles Street
Baltimore, MD 21218, USA
andreou@jhu.edu

SUSAN L. DENHAM
Cognition Institute, Plymouth University
Portland Square A219, Drake Circus
Plymouth, Devon PL4 8AA, UK
sdenham@plymouth.ac.uk

THOMAS WENNEKERS
Cognition Institute, Plymouth University
Portland Square A218, Drake Circus
Plymouth, Devon PL4 8AA, UK
thomas.wennekers@plymouth.ac.uk

Accepted 28 May 2013
Published Online 23 July 2013

The ability to recognize the behavior of individuals is of great interest in the general field of safety (e.g. building security, crowd control, transport analysis, independent living for the elderly). Here we report a new real-time acoustic system for human action and behavior recognition that integrates passive audio and active micro-Doppler sonar signatures over multiple time scales. The system architecture is based on a six-layer convolutional neural network, trained and evaluated using a dataset of 10 subjects performing seven different behaviors. Probabilistic combination of system output through time for each modality separately yields 94% (passive audio) and 91% (micro-Doppler sonar) correct behavior classification; probabilistic multimodal integration increases classification performance to 98%. This study supports the efficacy of micro-Doppler sonar systems in characterizing human actions, which can then

be efficiently classified using ConvNets. It also demonstrates that the integration of multiple sources of acoustic information can significantly improve the system's performance.

Keywords: Multimodal integration; sonar; convolutional network; human action recognition.

1. Introduction

Recognizing the behavior of individuals is important for understanding their state of mind and making behavioral decisions. There exist a wide range of human sensing methods for automatic behavior recognition, including cameras, contact and pressure sensors, thermal imagers, motion sensors, radars, and electric field sensors, among others.[1] Compared to other sensors, cameras are cheap, offer high spatial definition and provide a large amount of information regarding objects in the scene. Therefore, camera-based applications are commonly employed for automatic behavior recognition and numerous examples have been reported in the literature with strong emphasis on security applications such as road traffic,[2,3] or natural action classification.[4–7] However, the high-dimensionality of video-based sensor signals makes them more difficult to parse and commonly lead to a high number of false positives.[1] Additionally, conventional video-based systems for automatic behavior recognition have a number of drawbacks, such as requiring a lot of memory, computing power and communication bandwidth to process and transmit the images, being bulky and relatively immobile and raising privacy concerns when the system is deployed in public.

Here we present a fast, portable, low-cost and noninvasive alternative system that is capable of forming composite representations of behavior exclusively through the use of information derived from sounds. The system employs sound in two ways. First, it is able to identify and classify moving objects in the environment by active sound emission and the detection of micro-Doppler frequency shifts in sonar returns.[8] Second, it can identify the behavioral consequences of actions by passive detection and classification of sounds emitted by the objects themselves as they interact with their environment (e.g. foot-steps, hand-claps). By integrating these two modalities and forming associations between body movements (overt behavior) and sounds, this situated cognitive system goes beyond normal human capabilities and provides an acoustic analogy to a camera-based visual scene analysis system.

The micro-Doppler sonar system we employ is inspired by natural bio-sonar which echo-locating animals use to locate, range and identify objects.[9–11] Here we use a constant frequency signal that allows us to acquire signatures of motion from articulate objects, detected as modulations of the frequency in the emitted sonar signal; higher frequencies being generated by motion towards the device, and lower frequencies by movements away, the classic Doppler shift. This micro-Doppler sonar technology complements cameras and visual surveillance in situations where the mere presence of life is relevant (for security or search and rescue reasons, for example), since although it depends upon a clear line 'of sight' between the detector and the object of interest it does not rely on visibility per se. Preliminary work on human and animal species classification[12,13] as well as action and behavior recognition[14,15] yielded highly promising recognition rates. Here we present a novel approach that for the first time incorporates multi-scale analysis and multi-modal integration.

Convolutional Neural Networks (ConvNets)[16] are inspired by biophysiological principles derived from the study of primary visual cortex.[17] They provide a flexible and trainable hierarchical architecture that can learn selective and invariant features for classification. ConvNets have a long history and have been demonstrated to work robustly in character recognition as well as speech and time-series prediction.[18] Similar bio-inspired architectures have been used for movement[19] and face[20] recognition. The architectural simplicity of ConvNets proved to be suitable for custom hardware implementation[21] and a stream processor for ConvNets has also been built.[22]

A bio-inspired architecture for sound processing is employed in this paper. Although the mammalian auditory system is not fully understood, it has been shown that in monkeys it contains at least 20 interconnected regions.[23] The same study showed the auditory cortex of monkeys is hierarchically

organized with at least four distinct levels of processing. In this line, a recent fMRI study[24] demonstrated that the functional organization of the anterolateral processing pathway in humans is largely homologous to that of nonhuman primates, which has been hypothesized to mediate sound classification similar to the role of the ventral pathway in the visual system. Further support for a computational model of auditory object processing based on spectro-temporal features, is found in a study[25] that employed Dynamic Causal Modeling to evaluate the functional connectivity between three different regions of the auditory system. Regarding micro-Doppler sonar processing, evidence suggests that the auditory system in bats is also organized hierarchically.[26] All in all, convergent evidence suggests that ConvNets may capture some of the principles of auditory processing in cortex, such as the use of spectro-temporal features and a hierarchical organization.

Here we propose using ConvNets to model both passive sound and micro-Doppler sonar processing systems with application to behavior classification. Only a few previous studies have employed ConvNets to model auditory processing[27–30] (see the Discussion for a comparison) and none have used them to model micro-Doppler sonar processing. We test the action classification capabilities of our model on a multimodal dataset containing auditory and ultrasonic recordings of 10 people performing seven different actions. Additionally, we show that by probabilistically integrating the model output over time the classification accuracy on the individual modalities increases to 94% and 91% for passive sound and micro-Doppler, respectively. Further improvement is achieved through the multimodal integration of the passive auditory and the active micro-Doppler signatures, yielding 98% classification performance.

2. Methods

2.1. *Classifier architecture*

The classification system is based on ConvNets, which allow flexible architectures composed of several modules or building blocks that can be arranged differently according to the task requirements. The main building blocks, using a similar nomenclature and definition to that proposed by LeCun,[16] are described below. For each building block the variable x denotes the input 3D array composed of a_f 2D feature maps of size $a_1 \times a_2$; and y denotes the output 3D array composed of b_f 2D feature maps of size $b_1 \times b_2$. The ith feature map is denoted as x_i, and a specific component at location (j, k) within that feature map is denoted as x_{ijk}.

Filter Bank Convolution: Denoted as $n.F_{CTA}^{f_1,f_2}$, where n is the number of filters; f_1, f_2 define the filter size; and CTA represents the use of the convolution (C) operation, the tan$h$ (T) nonlinearity and the absolute (A) value function. This module computes the convolution of the input signal with a set of pre-learned filters and applies the tan$h$ and absolute function nonlinearities to the output, according to the following equation:

$$y_j = \left| \tanh\left(\sum_i w_{ij} * x_i \right) \right|, \tag{1}$$

where tan$h$ is the hyperbolic tangent nonlinearity; $*$ is the 2D discrete convolution operator; w_{ij} is a filter of size $f_1 \times f_2$ that connects input feature map x_i to output feature map y_j. Taking into account the border effects, the size of the output feature maps is: $b_1 = a_1 - f_1 + 1$ and $b_2 = a_2 - f_2 + 1$.

Local Contrast Normalization: Denoted as N_r, where r is the radius of a Gaussian window. This module performs local substractive and divisive normalization enforcing competition between the spatially surrounding units in the same feature map and between units coding different features at the same spatial location. The subtractive normalization operation for a given location x_{ijk} is implemented as:

$$v_{ijk} = x_{ijk} - \sum_{ipq} g_{pq} \cdot x_{i,j+p,k+q}. \tag{2}$$

v_{ijk} is the subtractively normalized output and g_{pq} a Gaussian window of radius r, sum-normalized to 1. Divisive normalization then computes:

$$y_{ijk} = v_{ijk} / \max(c, \sigma_{jk}), \tag{3}$$

$$\sigma_{jk}^2 = \left(\sum_{ipq} g_{pq} \cdot v_{i,j+p,k+q}^2 \right), \tag{4}$$

where y_{ijk} is the divisively normalized output; $c = \text{mean}(\sigma_{jk})$; and σ_{jk} is the weighted standard deviation of all features within the surrounding region of radius r. The normalization operation is inspired by computational models of primary visual cortex.[32]

Average Pooling and Subsampling is denoted as $P^{f_1,f_2}_{s_1,s_2}$, where f_1, f_2 are pooling sizes and s_1, s_2 are the subsampling step in each dimension:

$$y_{ijk} = \sum_{p=1\ldots f_1, q=1\ldots f_2} x_{i,j+p,k+q}/(f_1 \cdot f_2). \quad (5)$$

Each feature map y_i is then spatially subsampled by a factor of s_1 in the horizontal dimension and s_2, such that $b_1 = a_1/s_1$ and $b_2 = a_2/s_2$.

Classifier: This module employs the output of the previous layer to train a classifier using supervised learning and multinomial logistic regression or a linear support vector machine. In our simulations we employ the latter (implemented using the publicly available *libsvm* library[33]) and thus denote this top module as *SVM*. The SVM module is trained using labeled data from the layer immediately below, and generates multi-class probability distributions over the trained classes using pairwise coupling.[34]

2.2. *Signal pre-processing*

The only required signal pre-processing step is to compute a time-corrected instantaneous frequency reassigned (IFR) spectrogram representation of each recorded micro-Doppler signature, using the method of Nelson.[35] Prior to calculating the IFR spectrogram, the ultrasonic signal is mixed such that the emitter frequency is transformed to 2.5 kHz. The frequency range of the IFR for ultrasonic data is therefore set from 1.85 kHz to 3.15 kHz, i.e. centered around the down-sampled emitter or carrier frequency, in order to capture the frequency shifts generated by the moving body parts. Additionally, the IFR channels carrying information corresponding to the ultrasonic carrier frequency (2.5 kHz after down-sampling) are set to zero, in order to emphasize the micro-Doppler frequency shifts.

For the passive acoustic data the IFR frequency range is set between 100 Hz and 12 kHz, which approximately corresponds to the human audible frequency range. To eliminate a beep sound that signals the start of the trial, the IFR channels corresponding to this frequency are set to zero.

Other parameters of the IFR spectrogram function are the number of channels, which determines the frequency resolution (set to 50) and the bandwidth, which determines the number of samples or temporal resolution (set to 95). These values, derived in a previous study that used similar ultrasonic data,[15] provided a good compromise between reducing the size of the input and keeping enough information to maximize classification. However, we found that varying these values did not have a significant effect on performance, as long as the rest of the ConvNet parameters were tuned afterwards.

2.3. *Processing micro-Doppler signatures using ConvNets*

The micro-Doppler data was processed in a ConvNet composed of six layers, that receives input from the IFR spectrogram of the recorded micro-Doppler signal. The ConvNet has three pooling layers, two convolution layers and a classifier, see Fig. 1. Only the convolution and pooling modules constitute independent model layers, whereas the normalization module is included within the convolution layer. The size of the output of each layer, which takes into account the border effect of the convolution operation (the pooling operation has no border effect), is also included in Fig. 1. The configuration and parameters of the ConvNet were chosen as a trade-off between maximizing classification performance and minimizing simulation time. The parameters are shown in Table 1 and the Results section includes a robustness analysis for most of them. Following the nomenclature from the previous section, a network with selected parameters can be written in short form as:

$$P^{1,30}_{1,20} \rightarrow 8.F^{10,6}_{CTA}/N_3 \rightarrow P^{4,4}_{1,2} \rightarrow 64.F^{4,4}_{CTA}/N_0 \rightarrow P^{4,8}_{2,2} \rightarrow \text{SVM}. \quad (6)$$

The ConvNet can process an IFR spectrogram of any size and will generate a fixed number of output probability distributions corresponding to different time steps. Given the parameters of the network, the minimum number of spectrogram frames to generate an output probability distribution at the top layer is 530 (for the ultrasonic data). This corresponds to 2.57 s input signal. All receptive field sizes are shown in Table 1. These have been calculated taking into account the large overlap in the receptive fields (e.g. 10/30 for Layer 1). The high overlap in the network leads to a temporal step between top layer events of only 80 spectrogram frames (0.39 s of signal). Thus, for the example shown in Fig. 1, with an input spectrogram of 1025 frames (5 s of signal), the number of

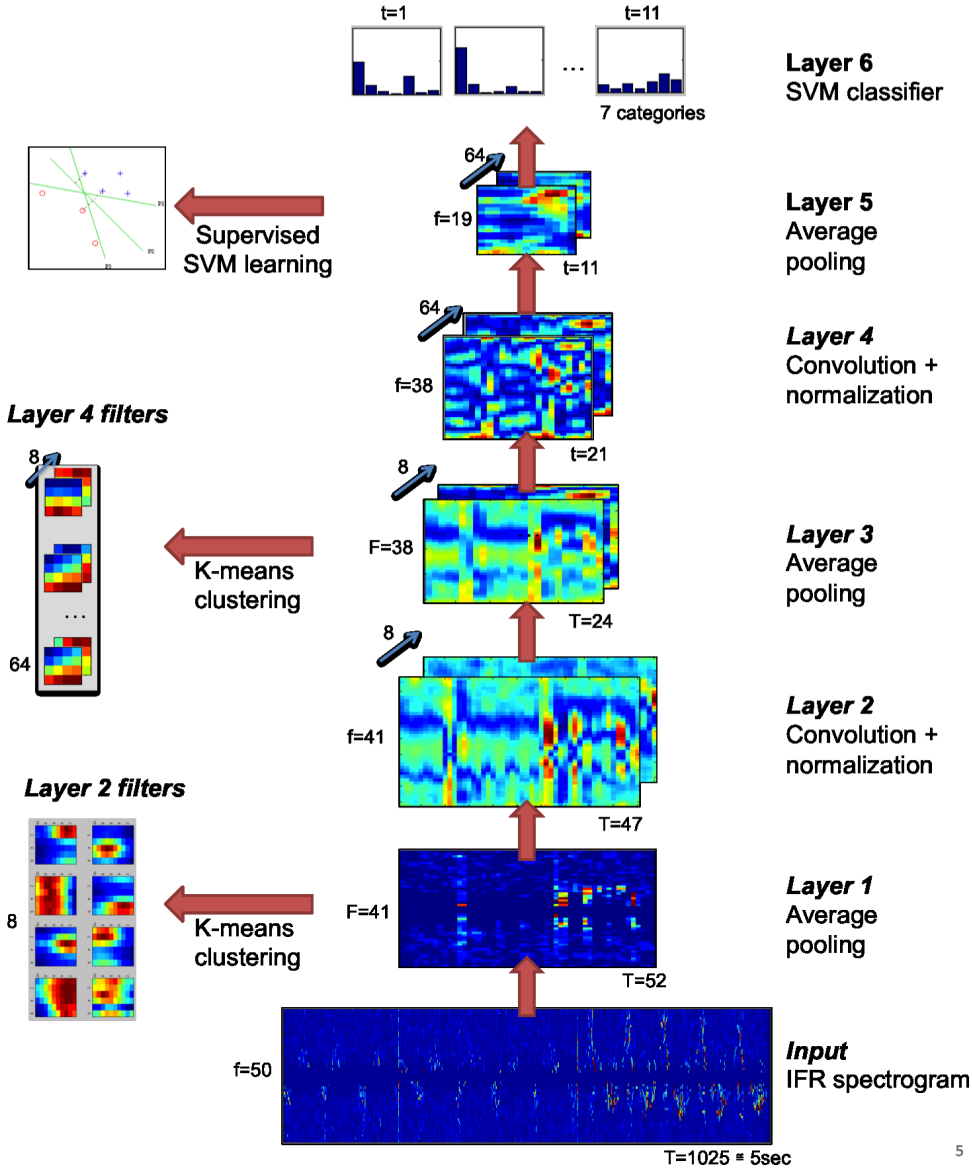

Fig. 1. Convolutional Neural Network for ultrasonic data classification. The output of each of the six layers in the network is shown schematically for an input signal of 5 s. The input is first pre-processed by calculating its IFR spectrogram, which serves as input to the model. Along the left-hand side, the different learning stages and resulting filters are also represented schematically. The output of the model consists of a probability distribution over the classes for each top layer time step, obtained using an SVM classifier.

output probability distributions at the top classifier layer should be $5/0.39 = 12.8$, but, due to the border effect of the convolution layers, the resulting number of output events is 11.

2.4. *Processing auditory data using ConvNets*

The ConvNet architecture used to process the auditory data has exactly the same parameters as that for the micro-Doppler data except for the first pooling layer, which is responsible for the temporal subsampling of the IFR spectrogram. Because the microphone sampling frequency was higher than that of the micro-Doppler device and the recording times of both devices were not always identical, the spectrograms corresponding to each modality differed in length. To obtain model outputs synchronized for both auditory and micro-Doppler data, we calculated

Table 1. Parameters of the ConvNet for ultrasonic (middle) and auditory (right) data processing.

Parameter	Ultrasonic	Auditory
	***Layer* 1**	
Type	Average pooling	
Pooling size (freq), f_1	1	1
Pooling size (time), f_2	30 (0.15 s)	72 (0.15 s)
Step size (freq), s_1	1	1
Step size (time), s_2	20 (0.10 s)	48 (0.10 s)
	***Layer* 2**	
Type	Convolution and normalization	
Number of filters, n	8	
Filter size (freq), f_1	10	
Filter size (time), f_2	6 (0.63 s)	
Normalization radius, r	3	
	***Layer* 3**	
Type	Average pooling	
Pooling size (freq), f_1	4	
Pooling size (time), f_2	4 (0.92 s)	
Step size (freq), s_1	1	
Step size (time), s_2	2 (0.19 s)	
	***Layer* 4**	
Type	Convolution and normalization	
Number of filters, n	64	
Filter size (freq), f_1	4	
Filter size (time), f_2	4 (1.21 s)	
Normalization radius, r	0	
	***Layer* 5**	
Type	Average pooling	
Pooling size (freq), f_1	4	
Pooling size (time), f_2	8 (2.57 s)	
Step size (freq), s_1	2	
Step size (time), s_2	2 (0.39 s)	
	***Layer* 6**	
Type	SVM linear classifier (*libsvm*)	
SVM options	$t = 0$ (linear kernel), $c = 1$ (cost), $b = 1$ (probability estimates)	

how much longer the auditory spectrogram was compared to the ultrasonic spectrogram and obtained an average scaling factor of 2.4. We use this value to modulate the pooling parameters of the auditory ConvNets such that the output of Layer 1 has the same length as for the ultrasonic ConvNets. The resulting architecture and parameters of the ConvNets for auditory data are described in Table 1 and indicated below using the short form notation:

$$P_{1,48}^{1,72} \rightarrow 8.F_{CTA}^{10,6}/N_3 \rightarrow P_{1,2}^{4,4} \rightarrow 64.F_{CTA}^{4,4}/N_0 \rightarrow P_{2,2}^{4,8} \rightarrow \text{SVM}. \quad (7)$$

2.5. *Learning using ConvNets*

The collected dataset is randomly divided into a training set with 50% of the subjects and a testing set with the other 50%. Thus, each set is composed of 175 files of approximately 10 s, corresponding to 7 actions × 5 subjects × 5 trials. The random selection process of subjects is repeated 50 times and the results are averaged for cross-validation.

To calculate the filters of the convolution layers we employ k-means clustering. We first compute from the previous layer's output all 3D patches of size $f_0 \times f_1 \times f_2$, where f_1 and f_2 are the spatial dimensions (i.e. the frequency and time dimension of the spectrogram) and f_0 is the number of feature maps in the previous layer. These patches are fed into the k-means clustering algorithm to obtain n cluster centers, which are used as filters after sum-normalizing each of them to one. Given that k-means is quite sensitive to initial starting conditions, we implemented a procedure for computing a refined starting condition that leads to improved solutions.[36]

The learning process is illustrated in Fig. 1, which includes examples of the filters learned by the micro-Doppler processing ConvNets. In this case, the 8 Layer 2 filters of size $f_0 = 1 \times f_1 = 10 \times f_2 = 6$ are learned from the output of Layer 1. Note that $f_0 = 1$ because the number of feature maps in Layer 1 is just one, i.e. the pooled IFR spectrogram. The 8 filters learned show similar characteristics to the spectro-temporal receptive fields found in primary auditory cortex.[37] The output of Layer 1 is then convolved with each of these filters to obtain the 8 output feature maps of Layer 2. Similarly, the 64 Layer 4 filters of size $f_0 = 8 \times f_1 = 4 \times f_2 = 4$ are learned from the output of Layer 3. In this case $f_0 = 8$ because Layer 3 is composed of eight feature maps. The output of Layer 3 is then convolved with each of the 64 learned filters, as described by Eq. (1), yielding the 64 feature maps of Layer 5.

To learn the linear SVM coefficients of the Layer 6 classifier we employ the *libsvm* library.[33] The output of Layer 5, x, has dimensions $a_f \times a_1 \times a_2$, where

a_f is the number of feature maps, a_1 represents the frequency dimension, and a_2 the temporal dimension. We then compute the vector of data points for the SVM, where each data point corresponds to one time step of x such that there are a_2 data points of size $a_f \times a_1$. As learning is supervised at this stage, we feed the SVM training function with a vector containing the class corresponding to each of the a_2 data points. We use the default cost value, $C = 1$, and activate the probability estimate option to allow for multi-class probability outputs during testing.

2.6. *Combining probability distributions over time*

We propose combining the output probability distributions of the ConvNet, $P(C|X_t)$, over several time steps, where X_t represents the input feature vector to Layer 6 at time step t, and $P(C|X_t)$ is the class probability given X_t. Assuming conditional independence of the input feature vectors given the class, i.e. $P(X_t, X_{t-1}|C) = P(X_t|C) \cdot P(X_{t-1}|C)$, and a uniform prior distribution, $P(C)$, we can write:

$$P(C|X_{t-n_{\text{win}}}, \ldots, X_t) = \alpha \cdot \prod_{d=1\cdots n_{\text{win}}} P(C|X_{t-d}), \tag{8}$$

where α is a normalizing constant that ensures that $\sum_k P(C_k|X_{t-n_{\text{win}}}, \ldots, X_t) = 1$; n_{win} is the window size and represents the number of probability distributions combined over time (in Layer 6 time steps); and $P(C|X_{t-n_{\text{win}}}, \ldots, X_t)$ represents the posterior distribution over classes given the input feature vectors between time steps $t - n_{\text{win}}$ and t.

2.7. *Multimodal integration*

To test whether combining information from both the micro-Doppler and auditory modalities improves the results we use a method of multimodal integration. Let $X_{t,u}$ and $X_{t,a}$ be the input feature vectors to Layer 6 at time step t for the ultrasonic and auditory ConvNets, respectively. Importantly, these outputs, $P(C|X_{t,u})$ and $P(C|X_{t,a})$, are aligned in time thanks to adapting the Layer 1 pooling parameters. Assuming conditional independence of $X_{t,u}$ and $X_{t,a}$ given the class, and a flat prior probability, $P(C)$, we can combine both sources of information using

$$P(C|X_{t,u}, X_{t,a}) = \alpha \cdot P(C|X_{t,u}) \cdot P(C|X_{t,a}), \tag{9}$$

where α is a normalizing constant that ensures that $\sum_k P(C|X_{t,u}, X_{t,a}) = 1$. $P(C|X_{t,u}, X_{t,a})$ represents the posterior probability over classes given the input feature vectors at time step t for both modalities.

2.8. *Real-time processing*

Real-time processing is accomplished by employing a C application that was developed to continuously read data blocks of fixed size (in our case, 19,400 bytes) from the micro-Doppler device and feed them to Matlab via a pipe such that no data is lost while other operations are carried out in real time. To process the incoming blocks of data, first the IFR spectrogram is calculated and then passed through the six-layer ConvNet that outputs a probability distribution over the classes. Given that the model operations have been efficiently implemented using Matlab's Image Processing toolbox, the ConvNet requires just 0.04 s of computation time for each second of data. Additionally, the receptive field size in the top layer was reduced from 8 to 4 such that the system only requires 1.8 s of data to provide the first output probability distribution, and then updates it every 0.4 s.

The real-time demo can be run in three different modes: (1) *learning*, to record the IFR spectrogram of a micro-Doppler signal, (2) *training*, to use stored IFR spectrograms for each action class to train the ConvNet model, and (3) *testing*, to compute the output probability distribution over actions classes for a given real-time input signal.

A graphical user interface was developed to visualize the input IFR spectrogram, output of the model layers, and the output probability distribution (http://scandle.eu/publications/posters-and-papers/fig_demo.png/view). The GUI allows the user to record new classes and train the model parameters to newly recorded data. Training is done offline, but very fast. This allows users to experiment with different sets of actions. For example, training the six-layer model with 60 s recordings of four different actions takes approximately three seconds.

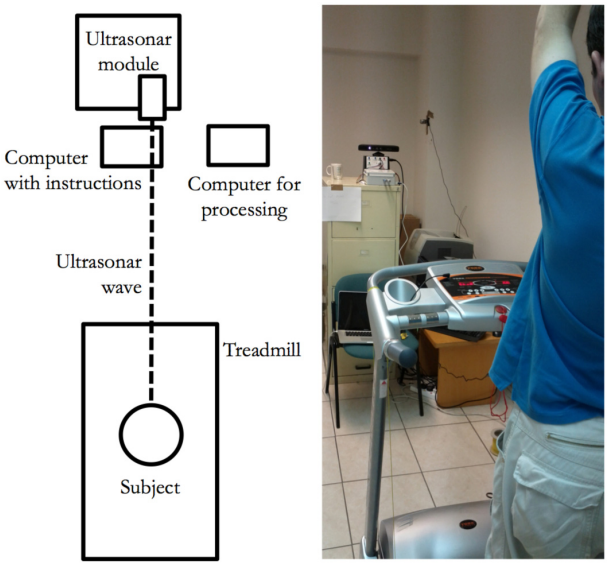

Fig. 2. Experimental setup for the collection of the multimodal dataset.

2.9. *Data collection*

A human behavior dataset was collected in an experimental indoor setup, see Fig. 2.

All actions were performed on top of a treadmill. The micro-Doppler sonar module was placed at approximately 2.5 m in front of the subject and at approximately 1.5 m above the floor. A treadmill, adjusted to minimize the surface area occluding the subject, was used to eliminate the effect of distance-modulated amplitude that is characteristic of the micro-Doppler signature. Without the treadmill, the amplitude of the recorded signal varies continuously as the subject moves closer to the micro-Doppler sonar. Additionally, the micro-Doppler signature would reflect a shift in its central frequency proportional to the velocity of the moving subject's body. The treadmill eliminates this highly salient cue, which would allow for trivial classification based on the average body speed, and action classification is based solely on the individual speeds of the moving body parts. The treadmill also facilitates recording actions such as walking or running given the limited distance range of the micro-Doppler sonar module (approximately 5 m). In previous experiments[31] these type of actions could only be recorded for short periods of time (1 or 2 s) and the sonar device had to be placed with a diagonal angle of incidence relative to the subject in motion, resulting in a reduced visibility of the body parts. The current setup solves these problems and provides a larger and cleaner dataset that facilitates the understanding and characterization of this novel sensory technology.

The setup also includes a studio quality microphone used to record the sounds, and the Kinect RGBD camera for ground truth. Data was collected simultaneously from the three sensors using a single graphical user interface (GUI) written in Matlab©.

Ten subjects (eight males and two females) performed five trials of seven different actions: (1) walking slowly (3 km/h), (2) walking fast (6 km/h), (3) running slowly (9 km/h), (4) running fast (12 km/h), (5) clapping hands, (6) calling "help me" while moving their arms up in the air and (7) calling "come here" while beckoning with their right arm. The total number of action trials was therefore 350, i.e. 50 for each action class. Each trial recording lasted approximately 10 s; it started after the subject had already begun performing the action and stopped before the subject had finished, so that the whole 10 s contained information related to the action. An example of a

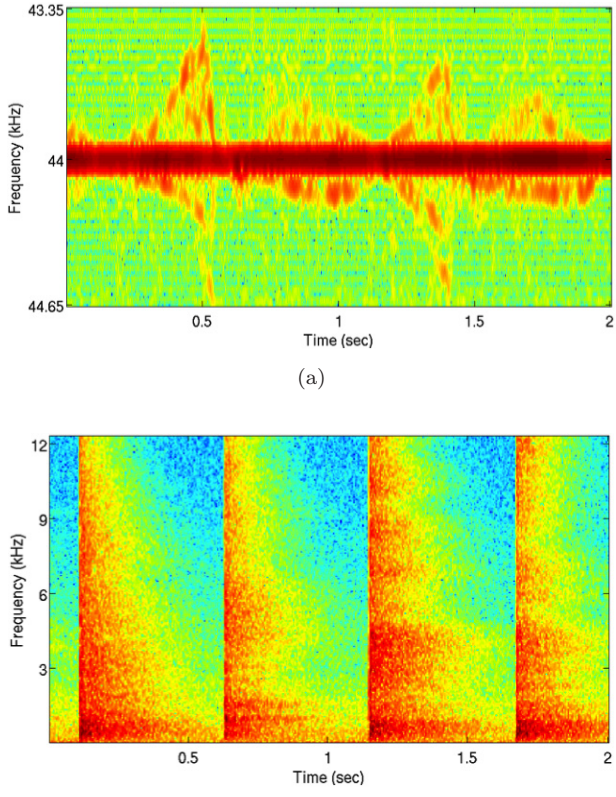

Fig. 3. Example spectrogram for the "clapping action" (a) micro-Doppler sonar signature and (b) microphone sound signal.

signature spectrogram of the micro-Doppler sonar and the microphone signals for the "clapping" action is shown in Fig. 3.

3. Results

3.1. *Classification performance and multimodal integration*

Figure 4 shows the classification performance of the ConvNets for micro-Doppler and auditory data separately, as well as their combined performance after integration. Results are shown as an average over all classes and for three individual classes: "clapping", "come here" and "fast running". The error bars indicate the standard deviation over the 50 repetitions using different random subsets of subjects for training and testing. Results are averaged over all classes and plotted as a function of the temporal integration window, n_{win}, where a window of 1 time step corresponds to approximately 2.57 s of input signal and a window of 13 time steps corresponds to approximately 7.25 s of input signal.

The average classification accuracy for micro-Doppler data ranges between 83.4% ($n_{\text{win}} = 1$) and 91.6% ($n_{\text{win}} = 13$); for auditory data between 88.2% ($n_{\text{win}} = 1$) and 92.2% ($n_{\text{win}} = 13$); and for multimodal integration between 95.3% ($n_{\text{win}} = 1$) and 97.6% ($n_{\text{win}} = 13$). With respect to individual classes, assuming $n_{\text{win}} = 13$, the classification accuracy for micro-Doppler data is lowest for the "come here" condition (83.0%) and highest for the "fast run" condition (97.4%); for auditory data the lowest accuracy is also for the "come here" condition (78.2%) and the highest for the "slow walk" (97.1%)

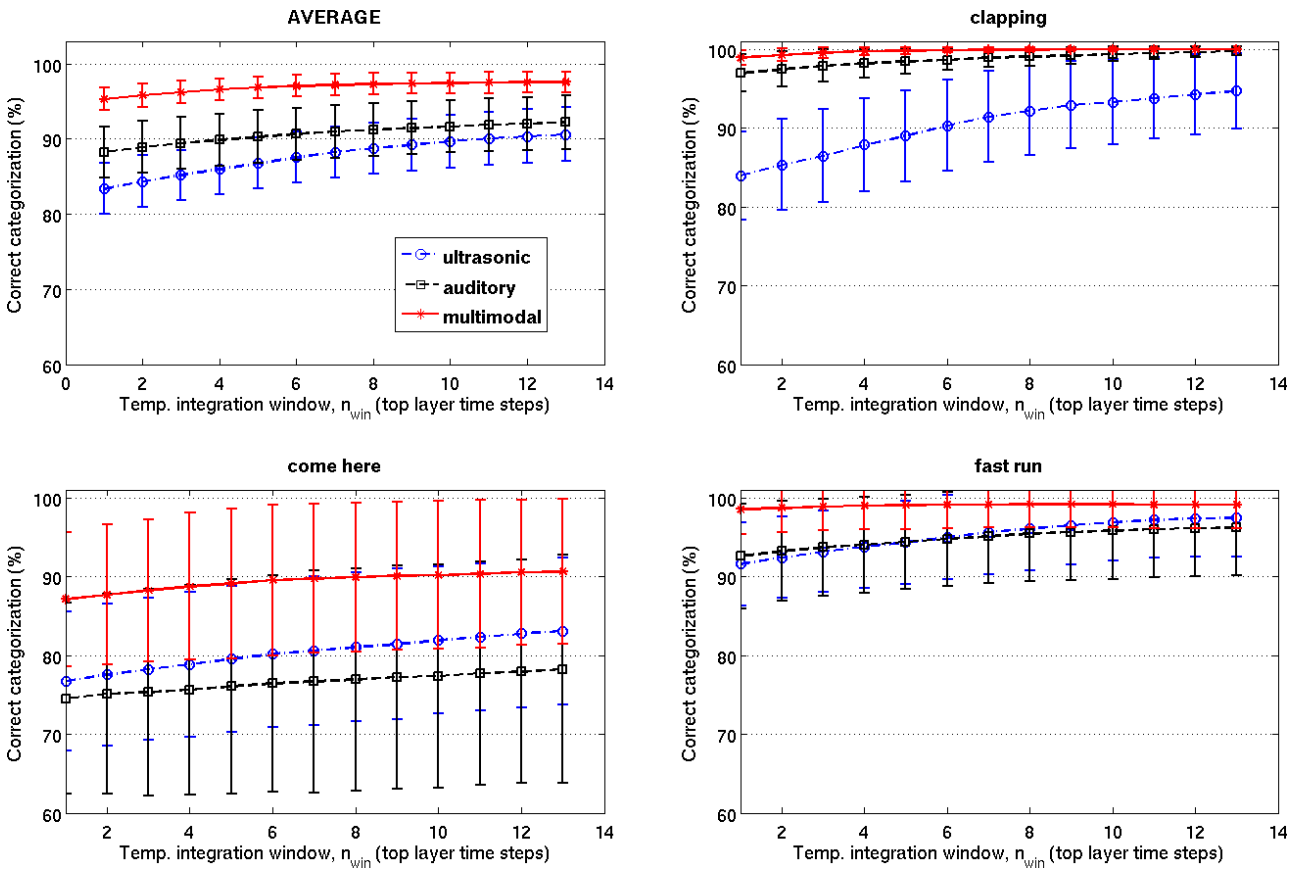

Fig. 4. Classification results for the ConvNet models of ultrasonic and auditory processing, as well as for multimodal integration. Results are shown for the overall average over classes and for three individual classes. The error bars indicate the standard deviation over the 50 repetitions using different random subsets of subjects for training and testing. The first top layer time step represents 2.57 s of data and each additional time step adds 0.39 s of new data, such that 13 time steps represent 7.25 s.

condition. The multimodal integration results, for $n_{win} = 13$, show an accuracy above 90% for all conditions, reaching almost perfect categorization for the "fast run" (99.1%), "fast walk" (99.3%) and "slow walk" (100.0%) conditions.

Interestingly the "come here" condition is improved substantially by multimodal integration even though the component modalities are rather poor in this case. Furthermore, the results when integrating modalities are always greater than those for each of the individual modalities; i.e. the combination always improves the individual modality performance.

3.2. *Comparison with classification performance of a Gaussian mixture model*

For comparison purposes, the same set of data was processed using a standard auditory classification method: computing a set of features based on the Mel-frequency cepstral coefficients (MFCC) of the signal, building a Gaussian Mixture Model for each class and using a Bayesian classifier. The parameters of the model were based on those from previous studies[13,38,39] but were optimized for this dataset, by tuning them independently for each modality. The frequency range of the MFCC was set to 1.5–3.5 kHz (ultrasonic) and 0–15 kHz (auditory), and the number of filters in the MFCC filterbank was set to 128 filters (ultrasonic) and 512 filters (auditory). The remaining parameters, i.e. number of dimensions in the MFCC feature vector (20, 40, 80 or 120), number of frames per 10-s file (10, 20 or 30) and the number of mixtures in the GMM (5, 10 or 20) were also tuned independently for each modality but resulted in the same optimum values for both, i.e. 80, 20 and 10, respectively.

The 80-dimensional feature vector was composed of the first 40 MFCC and the 40 first-order differential MFCC, in order to include temporal information. Consistent with previous studies,[13] reducing the number of dimensions to 20 or 40 reduced the performance, but increasing it to 120, did not show significant improvements. The feature vector was computed for each frame of the signal, where each 10-s file had 20 frames, similar to the Layer 5 output of the ConvNet. The parameters of the Gaussian mixture models were initialized using the k-means clustering algorithm (with initial clusters from centroids and variances of random partitions of the dataset) and learned using the expectation-maximization (EM) algorithm. A simple Bayesian classifier was used to compute the probability that the input frames from the test dataset belonged to each of the seven classes. The training and testing datasets and the random subsets of subjects were kept the same as for the ConvNet experiments. The classification results are summarized in Table 2.

Table 2. Classification performance of GMM versus ConvNet.

Classification (%)	Micro-Doppler	Auditory
MFCC + GMM	68.9 ± 2.9	86.0 ± 1.8
IFR + ConvNet	83.4 ± 3.1	88.2 ± 3.1

3.3. *Model robustness*

To study the robustness of the model to variations in parameters we computed the classification accuracy, with $n_{win} = 1$, for different values of these parameters. The analysis was performed only on the micro-Doppler data as this was the most challenging modality and, thus, the ConvNets parameters were mainly tuned to optimize its performance. The results reflect the average accuracy over 10 repetitions of random subsets of subjects, 50% for training and 50% testing. The number of repetitions was 10 instead of 50 in order to reduce the simulation time required to test the 16 different parameters analyzed. These parameters were: temporal and frequency pooling size of Layers 1, 3 and 5; temporal and frequency step size of Layers 1, 3 and 5; filter size (time and frequency) of Layers 2 and 4; and the number of filters in Layers 2 and 4. Figure 5 shows 2D contour plots of the categorization accuracy as a function of eight of these parameters. The range of values for each parameter was selected in order to provide a reasonable model output based on preliminary results. Thus, the X- and Y-axes do not necessarily have a linear scale. The scale bar of each graph indicates the minimum and maximum correct classification percentage, which measures the model's robustness to variations of those parameters.

Out of the 16 parameters analyzed, 14 exhibit a maximum classification accuracy variation of 10–15% (typically between 70% and 80%), for the range of values tested. The two exceptions are the Layer 3

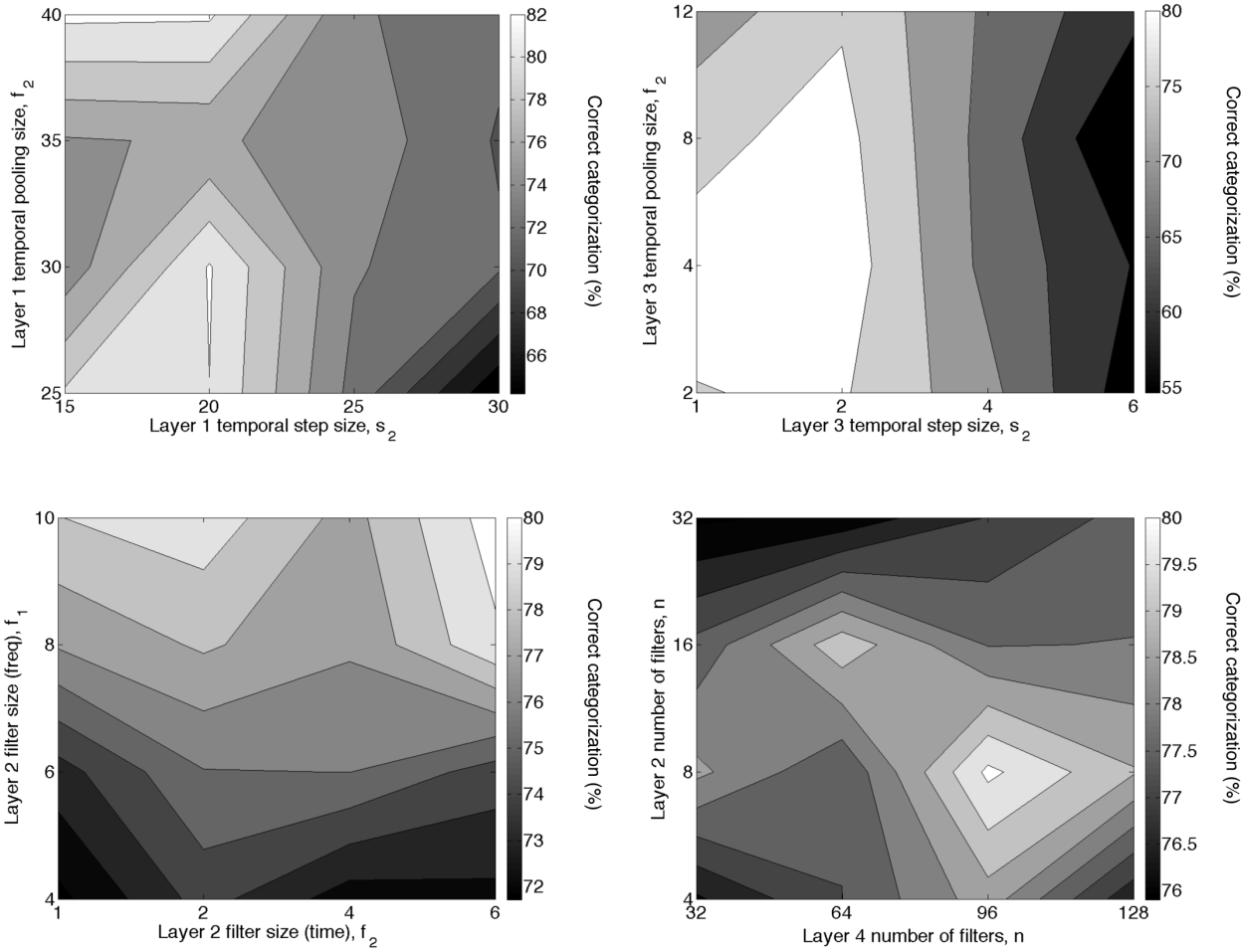

Fig. 5. Classification performance of the micro-Doppler model as a function of the ConvNet architecture parameters. The Figure contains 4 filled 2D contour plots of the correct classification percentage as a function of the following parameters: temporal pooling size of Layers 1 and 3; temporal step size of Layers 1 and 3; filter size (time and frequency) of Layer 2; and number of filters in Layers 2 and 4. The scale bar of each graph indicates the minimum and maximum correct classification percentage, providing a measure of the model robustness to variations of those specific parameters.

frequency step size, s_1, reaching an accuracy below 20% for values above 4; and the Layer 3 step size, s_2, reaching an accuracy below 60% for values above 6.

4. Discussion

4.1. *Comparison with existing auditory models based on ConvNets*

Despite the recent resurgence and success of ConvNets for object recognition, only a few studies have applied this methodology to the auditory domain. Here we analyze four of these models by comparing their input, architecture parameters and learning methods as shown in Table 3. The model proposed here can be applied to the classification of any sound, wheras previous models have focused on speech detection[27] or music classification.[28–30] All models employ some kind of spectro-temporal representation of the sound signal, applied over frames of a fixed length and typically with some overlap.

With respect to architecture, previous models have either two[27–29] or three[30] feature convolution layers, and only one of them[30] has separate pooling/subsampling layers. The rest[27–29] include a temporal subsampling step within the feature convolution layer. This means more information is lost in the temporal dimension as compared to models, such as the one proposed here, which by averaging over neighboring units before subsampling, preserve some of the contextual information. Additionally, we also

Table 3. Comparison of existing auditory models based on ConvNets.

Model	Input	Architecture parameters	Learning
Speech detection[27]	20 features (log SNR of mel-frequency bands) per 16 ms-frame	$25.F^{20,1} \rightarrow 25.F^{1,3}_{1,7} \rightarrow$ classifier	Stochastic gradient descent
Music classification[28]	13 features (MFCCs) per 190 frames of 27 ms (50% overlap)	$3.F^{10,13}_{1,4} \rightarrow 15.F^{10,1}_{1,4} \rightarrow 65.F^{10,1}_{1,4} \rightarrow$ classifier	Stochastic gradient descent
Music classification[29]	96 features (constant-Q transform) per 46 ms-frame (50% overlap)	$N_r \rightarrow 512.F^{96,1}$ frame or $4 \times 128.F^{24,1}$ octave $\rightarrow$ SVM	Predictive sparse decomposition
Music classification[30]	2 parallel streams: 12 chroma features (pitch) and 12 timbre features per beat	$2 \times 100.F^{12,8}$ beats $\rightarrow$ $P^{1,4}$ beats $\rightarrow 100.F^{200,8}$ bars $\rightarrow$ $P^{1,4}$ bars $\rightarrow$ logistic regression	Unsupervised feature learning with CDBN and Stochastic Gradient Descent

average pool and subsample the frequency domain in order to increase invariance to transformations and distortions in that domain.

Previous learning algorithms include unsupervised methods that learn features useful for classification, such as Predictive Sparse Decomposition[29] or Contrastive Divergence using an equivalent Convolutional Deep Belief Network (CDBN)[30]; and supervised methods, such as Stochastic Gradient Descent[27,28,30], that learn to associate the top layer output to the different classes and refine the intermediate layer features. In this paper, we employ k-means clustering to learn the intermediate layer features in an unsupervised manner, and use supervised learning to train a linear SVM that can assign the Layer 5 output to different classes.

Regarding ultrasonic processing, this is the first model, to the best of our knowledge, that applies ConvNets to the classification of micro-Doppler signatures obtained using an ultrasonic device.

4.2. *Classification performance and multimodal integration*

The classification accuracy of the ConvNet models increases with the size of the cumulative probability window. For the ultrasonic data it ranges from 84% to 91%; for the auditory data, from 88% to 94%; and for multimodal integration, from 95% to 98%. This outperforms previous models on the same[14,15] and similar[31,40] ultrasonic data, and constitutes a novel benchmark for auditory and multimodal data.

Our results were compared to those obtained using MFCC features and a GMM, a methodology that has been proven to be effective for sound-based speaker identification[38] and walker identification[13,39] from micro-Doppler signatures. The classification performance was significantly lower for the ultrasonic data (65%) and slightly lower for the auditory data (80%) as compared to the ConvNet model. In both cases the same training and testing data subsets were employed and the 10-s files were segmented into the same number of frames or events before the classification stage. This suggests the hierarchical structure of the ConvNet might have advantages over the GMM in coping with the variability and generalizing to the new data.

Multimodal integration improves the results for all of the classes suggesting that the information contained in the ultrasonic and auditory modalities complement each other. For example, for the classes with a great variability in the sounds, i.e. the "come here" and "help me" classes, the ultrasonic data provides strong cues that compensate the decreased performance of the auditory model. Similarly, for classes such as "slow walking" where the ultrasonic performance is not so high, the auditory model can recognize the clear step sounds to compensate for this and achieve an overall high multimodal performance.

It is possible that the high classification results are partly due to the regularity in the step sounds and movements of the conditions that were

performed using the fixed speeds of the treadmill. However, this does not seem to be the main contributing factor given that for the "clapping" condition the rate and intervals of the claps was highly heterogeneous amongst subjects and the classification performance is still very high.

Interestingly, the model copes well with actions with short periodic cycles, such as "fast running", and those with longer periodic cycles, such as "help me". The hierarchical nature of the model implies that features learned at different layers will have different time scales. This suggests that for composite actions like the ultrasonic signature of "slow walking", lower level features may encode short components of the action, e.g. the forward movement of the right leg, whereas the higher level features may encode longer periods of the action, such as a complete walking cycle. Similar composite actions in the auditory domain include the "come here" and "help me" classes, which can be hierarchically organized into phonemes and words, for example. Notably, the ConvNet is also able to categorize actions with very short cycles, such as "clapping" or the walking and running conditions in the auditory domain, suggesting that the higher level features can also be composed of several periodic cycles of the same action.

The model results have been averaged over 50 different trials with randomized subject subsets for cross-validation. This prevents the model from being overfitted to the data and means the model is able to generalize to new subjects. The error bars also indicate that there is a high variability across trials, especially for the "come here" and "help me" conditions. This is consistent with the experimental design, given that during these two conditions the subjects were free to decide the frequency and duration of the gestures performed and the accompanying utterances. Part of the variability is also due to initialization of k-means clusters. By running 50 trials with the same subject subset the average standard deviation due to k-means initialization was estimated to be 1.2% for the ultrasonic modality and 3.3% for the auditory modality.

For the combination of probability distributions over time and the integration of both modalities it is possible to use different methods as proposed here. For example, we tested the method of averaging the probability distributions, which yielded only slightly worse classification results than our joint probability distribution method: an average of 0.1% less for the combination of probability distributions over time and 0.9% less for multimodal integration.

4.3. *Robustness*

The results in Fig. 5 provide an indication of the robustness of the model to variations in the networks parameters. However, these are not intended to provide an exhaustive robustness analysis, as they are limited in a number of ways. First, the analysis was only performed for the ultrasonic model and not for the auditory model or the multimodal integration. Second, the results only take into account a single top layer time step, whereas the classification performance after integrating several time steps is likely to yield improved robustness patterns. Finally, the parameter space was analyzed by varying the values of two parameters at a time, while keeping the rest of parameters fixed.

Despite these constraints, the results suggest that the model is robust to moderate variations of most of the architecture parameters (pooling size, step size and number of features). Two exceptions are the temporal and frequency step sizes in Layer 3 (subsampling), where higher values drastically reduce the model performance. This is reasonable given that increasing the subsampling step is directly proportional to the amount of information lost, such that doubling the step size means losing 50% of the information. This can represent a considerable loss given that the input to the model has already been significantly subsampled in the frequency domain by the IFR spectrogram computation and in the temporal domain by the Layer 1 operations.

4.4. *Future work*

Future lines of research are intended to explore the applicability of the sensor to real-life scenarios. In this sense, experiments will be developed to evaluate aspects such as the distance limits of the system, especially in outdoor conditions, and the effects on accuracy of the angle of incidence between the ultrasonic module and the target object. One key aspect here is the potential active control of the micro-Doppler sonar for interrogating the scene: unlike audio which comes from all directions without control, the sonar device can be activated intermittently and directed towards the desired targets.

Another interesting extension would be the implementation of the whole model on a FPGA, in order to obtain a compact and fast system that can be run in real time. Recent encouraging studies[22] demonstrate the feasibility of running ConvNets on FPGAs and the excellent performance that can be achieved. Although this approach has been typically limited to vision systems, preliminary results from work still in progress shows the successful implementation on FPGA of the proposed ConvNet for ultrasonic and auditory action classification.

5. Conclusion

In this paper, we present a novel system that integrates a micro-Doppler sonar system, a data collection experiment and a ConvNet model that processes micro-Doppler and auditory data for human action and behavior classification. The biologically-inspired ConvNets model can use the micro-Doppler data for classification of human actions with high efficiency and can be coupled with a parallel ConvNet architecture for passive sound recognition to further improve its performance. The model shows robustness to parameter variations and runs in real time, as demonstrated by the demo already implemented. This is partly thanks to homogeneity of the network (weight sharing), which also enhances the generalization capacity of the model, even with limited training data. These properties make the system suitable for a wide range of applications in the context of security, surveillance and human behavior monitoring.

The main novelties of the proposed integrated system are, first, performing action classification using micro-Doppler sonar signals derived from a custom made compact hardware system of low power consumption and cost. Second, the use of ConvNets to process micro-Doppler ultrasonic data based on an efficient computational implementation that can be run in real time. And third, the probabilistic integration of two complementary sources of acoustic information, ultrasonic and auditory, that significantly improves the system's performance.

Acknowledgment

This research was supported by the European Community's Seventh Framework Programme, grant no. 231168-SCANDLE: acoustic SCene ANalysis for Detecting Living Entities (http://www.scandle.eu).

References

1. T. Teixeira, G. Dublon and A. Savvides, A survey of human-sensing: Methods for detecting presence, count, location, track, and identity, Tech. Rep. 09-2010, ENALAB, Yale University (2010).
2. M. Yguel, O. Aycard, D. Raulo and C. Laugier, Grid based fusion of off-board cameras, *Intelligent Vehicles Symp.* (2006), pp. 276–281.
3. T. Ellis, Multi-camera video surveillance, *Int. Carnahan Conf. Security Technology* (2002), pp. 228–233.
4. C. Wallraven, M. Schultze, B. Mohler, A. Vatakis and K. Pastra, The poeticon enacted scenario corpus a tool for human and computational experiments on action understanding, *Int. Conf. Automatic Face Gesture Recognition and Workshops* (*FG 2011*) (2011), pp. 484–491.
5. F. Brémond, M. Thonnat and M. Zúniga, Video-understanding framework for automatic behaviour recognition, *Behav. Res. Meth.* **38** (2006) 416–426.
6. J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper and R. Kasturi, Understanding transit scenes: A survey on human behavior-recognition algorithms, *IEEE Trans. Intell. Transp. Syst.* **11** (2010) 206–224.
7. K. Subramanian and S. Suresh, Human action recognition using meta-cognitive neuro-fuzzy inference system, *Int. J. Neural Syst.* **22** (2012) 1250028.
8. Z. Zhang, P. O. Pouliquen, A. Waxman and A. G. Andreou, Acoustic micro-Doppler radar for human gait imaging, *J. Acoust. Soc. Am.* **121** (2009) EL110–EL113.
9. H.-U. Schnitzler and A. Denzinger, Auditory fovea and Doppler shift compensation: Adaptations for flutter detection in echolocating bats using CF-FM signals, *J. Comp. Physiol. A* **197** (2011) 541–559.
10. R. Kober and H. Schnitzler, Information in sonar echoes of fluttering insects available for echolocating bats, *J. Acoust. Soc. Am.* **87** (1990) 882–896.
11. G. Neuweiler, How bats detect flying insects, *Phys. Today* **33** (1980) 34–40.
12. Z. Zhang, P. Pouliquen, A. Waxman and A. Andreou, Acoustic micro-doppler gait signatures of humans and animals, *Annual Conf. Information Sciences and Systems, CISS '07* (2007), pp. 627–630.
13. Z. Zhang and A. Andreou, Human identification experiments using acoustic micro-doppler signatures, *Argentine School of Micro-Nanoelectronics, Technology and Applications, EAMTA 2008* (2008), pp. 81–86.
14. G. Garreau, C. Andreou, A. Andreou, J. Georgiou, S. Dura-Bernal, T. Wennekers and S. Denham, Gait-based person and gender recognition using micro-doppler signatures, *Biomedical Circuits and Systems Conf.* (*BioCAS*) (2011), pp. 444–447.
15. S. Dura-Bernal, G. Garreau, C. M. Andreou, A. G. Andreou, J. Georgiou, T. Wennekers and S. L. Denham, Human action categorization using ultrasound

micro-doppler signatures, in *HBU* (Springer, 2011), pp. 18–28.
16. Y. LeCun, K. Kavukcuoglu and C. Farabet, Convolutional networks and applications in vision, in *Proc. 2010 IEEE Int. Symp. on Circuits and Systems* (*ISCAS*) (2010), pp. 253–256.
17. D. H. Hubel and T. N. Wiesel, Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat, *J. Neurophysiol.* **28** (1965) 229–289.
18. Y. LeCun and Y. Bengio, Convolutional networks for images, speech and time-series, *The Handbook of Brain Theory and Neural Networks*, ed. M. Arbib (MIT Press, 1995), pp. 255–258.
19. M. A. Giese and T. Poggio, Neural mechanisms for the recognition of biological movements, *Nat Rev. Neurosci.* **4** (2003) 179–192.
20. N. T. N. Wi, C. K. Loo and L. Chockalingam, Biologically inspired face recognition: Toward pose-invariance, *Int. J. Neural Syst.* **22** (2012) 1250029.
21. B. Boser, E. Sackinger, J. Bromley, Y. Le Cun and L. Jackel, An analog neural network processor with programmable topology, *IEEE J. Solid-State Circuits* **26** (1991) 2017–2025.
22. C. Farabet, Y. LeCun, K. Kavukcuoglu, E. Culurciello, B. Martini, P. Akselrod and S. Talay, Large-scale FPGA-based convolutional networks, in *Scaling up Machine Learning: Parallel and Distributed Approaches*, eds. R. Bekkerman, M. Bilenko and J. Langford (Cambridge University Press, 2011), pp. 399–419.
23. J. H. Kaas and T. A. Hackett, Subdivisions of auditory cortex and processing streams in primates, *Proc. Natl. Acad. Sci. U.S.A.* **97** (2000) 11793–11799.
24. M. Chevillet, M. Riesenhuber and J. P. Rauschecker, Functional correlates of the anterolateral processing hierarchy in human auditory cortex, *J. Neurosci.* **31** (2011) 9345–9352.
25. S. Kumar, K. E. Stephan, J. D. Warren, K. J. Friston and T. D. Griffiths, Hierarchical processing of auditory objects in humans, *PLoS Comput. Biol.* **3** (2007) e100.
26. N. Suga, Principles of auditory information-processing derived from neuroethology, *J. Exp. Biol.* **146** (1989) 277–286.
27. S. Sukittanon, A. C. Surendran, J. C. Platt and C. J. C. Burges, Convolutional networks for speech detection, *Interspeech'04* (2004), pp. 433–445.
28. T. L. H. Li, A. B. Chan and A. H. W. Chun, Automatic musical pattern feature extraction using convolutional neural network, in *Proc. Int. MultiConf. Engineers and Computer Scientists, IMECS*, Vol. I (2010), pp. 546–550.
29. M. Henaff, K. Jarrett, K. Kavukcuoglu and Y. LeCun, Unsupervised learning of sparse features for scalable audio classification, in *Proc. Int. Symp. Music Information Retrieval (ISMIR'11)*, eds. A. Klapuri and C. Leider (University of Miami, 2011), pp. 681–686.
30. S. Dieleman, P. Brakel and B. Schrauwen, Audio-based music classification with a pretrained convolutional network, in *Proc. Int. Symp. Music Information Retrieval (ISMIR'11)*, eds. A. Klapuri and C. Leider (University of Miami, 2011), pp. 669–674.
31. G. Garreau, N. Nicolaou, C. Andreou, C. D'Urbal, G. Stuarts and J. Georgiou, Computationally efficient classification of human transport mode using micro-doppler signatures, *Annual Conf. Information Sciences and Systems* (*CISS*) (IEEE, 2011), pp. 1–4.
32. M. Carandini and D. J. Heeger, Normalization as a canonical neural computation, *Nat. Rev. Neurosci.* **13** (2012) 51–62.
33. C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.* **2** (2011) 27:1–27:27.
34. T.-F. Wu, C.-J. Lin and R. C. Weng, Probability estimates for multi-class classification by pairwise coupling, *J. Mach. Learn. Res.* **5** (2004) 975–1005.
35. S. A. Fulopa and K. Fitz, Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications, *J. Acoust. Soc. Am.* **119** (2006) 360–371.
36. P. S. Bradley and U. M. Fayyad, Refining initial points for K-means clustering, in *ICML '98: Proc. Fifteenth Int. Conf. Machine Learning*, San Francisco, CA, USA (Morgan Kaufmann Publishers Inc., 1998), pp. 91–99.
37. N. Mesgarani, S. V. David, J. B. Fritz and S. A. Shamma, Phoneme representation and classification in primary auditory cortex, *J. Acoust. Soc. Am.* **123** (2008) 899–909.
38. D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models, *IEEE Trans. Speech Audio Process.* **1** (1995) 72–83.
39. K. Kalgaonkar and B. Raj, Acoustic doppler sonar for gait recognition, *Advanced Video and Signal Based Surveillance, AVSS 2007* (2007), pp. 27–32.
40. K. Kalgaonkar and B. Raj, One-handed gesture recognition using ultrasonic doppler sonar, *Acoustics, Speech and Signal Processing, ICASSP 2009* (2009), pp. 1889–1892.