Multimodal Classification of Activities of Daily Living inside Smart Homes

Vit Libal[1], Bhuvana Ramabhadran[1], Nadia Mana[2], Fabio Pianesi[2], Paul Chippendale[2], Oswald Lanz[2], Gerasimos Potamianos[3⋆]

[1] IBM Thomas J. Watson Research Center, Yorktown Heights, New York, U.S.A.
[2] Fondazione Bruno Kessler (FBK), Trento, Italy
[3] Institute of Computer Science (ICS), FORTH, Heraklion, Greece

Emails: [1]{libalvit,bhuvana}@us.ibm.com; [3]gpotam@ieee.org;
[2]{mana,pianesi,chippendale,lanz}@fbk.eu

Abstract. Smart homes for the aging population have recently started attracting the attention of the research community. One of the problems of interest is this of monitoring the activities of daily living (ADLs) of the elderly aiming at their protection and well-being. In this work, we present our initial efforts to automatically recognize ADLs using multimodal input from audio-visual sensors. For this purpose, and as part of Integrated Project Netcarity, far-field microphones and cameras have been installed inside an apartment and used to collect a corpus of ADLs, acted by multiple subjects. The resulting data streams are processed to generate perception-based acoustic features, as well as human location coordinates that are employed as visual features. The extracted features are then presented to Gaussian mixture models for their classification into a set of predefined ADLs. Our experimental results show that both acoustic and visual features are useful in ADL classification, but performance of the latter deteriorates when subject tracking becomes inaccurate. Furthermore, joint audio-visual classification by simple concatenative feature fusion significantly outperforms both unimodal classifiers.

1 Introduction

Automatic recognition of human activities of daily living (ADLs) is deemed a crucial component of behavior pattern analysis systems. In the scope of the European-Union funded Netcarity Integrated Project [1], such a system would enable automatic assistive services for the elderly, providing the opportunity to increase their independence at the home environment. There are many examples of other potential applications of ADL recognition beyond elderly care, e.g. safety systems and services, security systems, situation-aware human-computer interfaces, etc. It is expected that the significance of ADL recognition technology will grow over time as sensor technology progresses and computational power increases.

⋆ G. Potamianos is currently with the Institute of Informatics & Telecommunications (IIT), National Centre of Scientific Research "Demokritos", Athens, Greece.

Among the various types of sensors that could be used to capture ADL information at home, we are interested in far-field microphones and cameras, due primarily to their unobtrusiveness and low cost. Using such sensors requires robust processing of the acquired signals, which constitutes a very challenging problem in realistic, unconstrained smart home environments. On the microphone side, for example, the task of ADL recognition becomes closely related to the field of acoustic scene analysis. The majority of previous research on that topic has mainly aimed at the recognition of short time-span, simple acoustic events in both smart office [2, 3] and smart home environments [4, 5]. ADLs, in contrast, constitute long time-span, complex activities.

In recent work [6], we have started to investigate the problem of ADL recognition using far-field microphones inside smart homes. There, the ADLs were modeled either as monolithic acoustic segments or as structured elements that can be decomposed into a sequence of shorter characteristic acoustic events. The former approach achieved the best results, when used in conjunction with an SVM classifier built on GMM supervectors [6]. However, this work did not take into consideration visual information, available through camera sensors. This could potentially impact performance, since the visual modality is generally known to help a number of perception technologies in smart spaces by complementing audio information [3]. Recently, for example, the visual modality has been successfully employed to recognize longer-span human activity in the office environment, jointly with audio information [7]. There, unconstrained realistic data was captured by one camera and microphone per room in a five-room environment. The classification used hidden Markov models and detected four basic office activities, namely "paperwork", "phone call", "meeting", and "Nobody in the Office".

Motivated by the above, in this paper we extend our prior work [6], by investigating the use of visual information in addition to far-field acoustic input for recognizing ADLs inside smart homes. For this purpose, we use ADL data acted by a number of subjects, recorded under realistic conditions using far-field microphones and cameras inside an apartment that has been set up as a smart home by Netcarity partner site FBK in Trento, Italy [8]. In this initial effort to incorporate visual information, we propose the use of human 3D location information as visual features, taking advantage of recent advances in multi-camera tracking [9]. Furthermore, we investigate the use of bimodal information, i.e. both acoustic and visual features, demonstrating significant performance improvements over both audio-only and visual-only systems, even though we utilize a relatively simple feature fusion approach. Note that since the emphasis of the paper is on extracting and incorporating visual information for ADL recognition, we limit ourselves to the use of relatively simple statistical methods – namely Gaussian mixture model classifiers. More complex modeling approaches could of course be used, leading potentially to further improvements [6, 7].

It is worth reiterating that the problem of ADL recognition in unconstrained realistic home environments using far-field audio-visual sensors is extremely challenging. For example, ADLs may overlap, multiple subjects may be present, and there typically exists significant variability in the background acoustic noise and

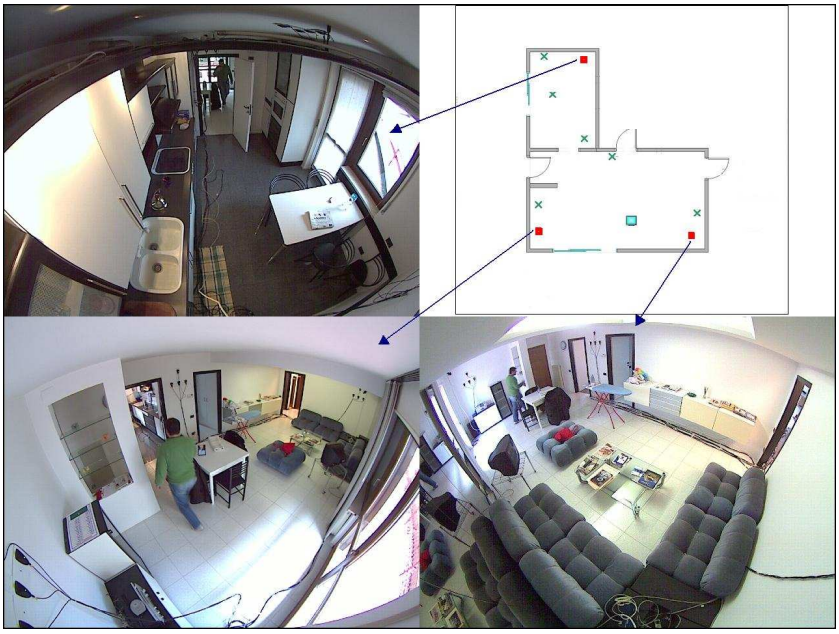

Fig. 1. A schematic diagram of the smart home used in the ADL corpus collection [8], depicting the audio-visual sensor locations (upper right). Example views of the three cameras are also shown.

changing spatio-temporal illumination conditions. In addition, ADLs are long and of complex structure, and they are typically characterized by "acoustically sparse" audio data with poor distinctive acoustic footprint. Therefore, and in order to simplify the complexity of the ADL recognition problem, we limit ourselves to a small set of six ADL classes, as discussed in Section 2. Furthermore, we assume that the temporal boundaries of ADLs are a-priori known, so the problem practically reduces to that of classification (instead of detection).

The remainder of the paper is structured as follows: Section 2 describes the data corpus, and Section 3 details our approach to ADL classification, including single-modality feature extraction and audio-visual fusion. Section 4 is devoted to the experiments, and finally Section 5 concludes with a summary of results and a brief discussion.

2 The Netcarity ADL Corpus

For our experiments we employ a specially designed audio-visual corpus of ADLs [8], collected as part of the Netcarity project [1]. The data was recorded in a real apartment, where two rooms – the living room and the kitchen – were equipped with a total of six T-shaped omni-directional microphone arrays with four microphones per array, thus providing 24 audio channels, as well as three

ADL	training set		test set	
class	# seg.	dur.	# seg.	dur.
EAT	128	83.44	32	82.34
RDG	128	66.06	32	71.22
TVW	128	97.24	32	97.52
IRN	64	96.85	16	98.04
CLN	64	65.38	16	66.66
PHN	64	104.49	16	103.62
total	576	84.42	144	85.67

Table 1. Data statistics for the training and test sets of the ADL corpus [8] used in the experiments in Section 4. Number of segments (# seg.) and their average duration (dur.) in seconds are depicted for each of the six ADL classes of interest.

cameras (two in the living room and one in the kitchen) with relatively wide fields of view. Each audio channel provided data at 16 bits per sample and a 48 kHz sampling rate, whereas the cameras yielded 640×480-pixel frames at approximately 10 frames per second. Fig. 1 depicts a schematic diagram of the apartment, with the microphone array and camera positions indicated by "x" marks and squares, respectively. Example video frames from the three cameras are also shown.

The collection was organized into 20 sessions, each about 1.5 hour long. Each session contained one main subject performing a prescribed set of 12 activities, randomly repeated four times. In order to obtain a realistic acoustic environment, three of these 12 activities were performed by the main subjects, while an interfering subject conducted other activities (e.g. the main subject could be watching TV, while the interfering subject was having a phone conversation). All 20 collected sessions are employed in our experiments: 16 sessions are used for training and 4 sessions for testing. Due to the fact that each subject appeared in one session only, this yields a speaker-independent training-testing scenario.

For the ADL classification experiments we limit ourselves to six classes: "eating-drinking" (EAT), "reading" (RDG), "ironing" (IRN), "cleaning" (CLN), "phone answering" (PHN), and "TV watching" (TVW). There exist 720 segments for these ADLs – 576 in training and 144 in the test set. Their detailed occurrence statistics in the two sets are depicted in Table 1.

3 Feature Extraction and ADL Classification

Gaussian mixture models (GMMs) are used to model the feature vector distribution for each ADL. In all audio-only, visual-only, and audio-visual classification, these vectors are available at a rate of 100 frames per second, and, during training, they are used to estimate six models (one per class) by means of the expectation-maximization algorithm. At testing, maximum-a-posteriori estimation is employed to determine the most likely ADL, assuming feature independence and equal ADL priors.

3.1 Acoustic Processing

To extract audio features, the signal from a single only microphone is used. For this purpose a centrally located microphone is selected, so that the average distance to the events happening in the apartment is minimal. From this signal, 13-dimensional perceptual linear prediction (PLP) coefficients with segment-level cepstral mean subtraction applied are used as acoustic feature vectors. Each ADL segment is represented by a sequence of PLP feature vectors, extracted at 10 Hz from a 25 ms Hamming-windowed signal, with 15 ms overlap between successive audio segments.

3.2 Visual Processing

In addition to audio signals, visual analysis of a scene can provide us with a rich set of features to detect ADLs. For the specific task of ADL recognition in a natural environment, we originally employed two separate detectors that generate both high-level (person tracking) and low-level (body activity) features. However, in this paper, we only utilize the positional information relating to the target's 3D location.

To detect position, a multi-camera particle filter tracker is used. At each filter update it generates a number of 3D position hypotheses for each target using the previous estimate and a simple motion model. For each new hypothesis a coarse 3-dimensional shape model is projected onto each of the calibrated camera frames, and color histograms are extracted from the identified image regions. The position hypothesis is then scored according to how well the extracted histograms match a previously acquired color model of the target. Position hypotheses with low scores are rejected, while those with high scores are maintained. The accuracy of the tracking depends greatly on the quality of the target model, which was acquired automatically as a new target entered the larger room (living room). More details can be found in [9].

Visual subject tracking provides the system with a location output at a variable frame rate, which on average is 10 frames per second. Before the features are used by the GMM classifier, they are upsampled to the same frequency as the audio, i.e 100 Hz, by means of linear interpolation. Due to difficulties experienced during the video acquisition process (chunks of missing frames and some non-synchronized sections) and strong changes in the color temperature of the lighting (sunlight and incandescent sources), the visual tracking of the subjects varied greatly in accuracy and consistency. In some situations the tracker was not able to detect the subject, mainly because the timestamps of the images delivered by the two cameras were misaligned, or the target showed low contrast with the background. Consequently, the tracking was not initialized and no positional data is available for ADL classification. Additionally, there is no subject tracking in the kitchen room, as only a single camera was available.

In order to assess the impact of the loss of visual tracking on the classification accuracy, for the visual modality, we designed five "nested" data sets for our experiments. These data sets were created from the original data by dropping

# Gaus.	Acc. (%)	# Gaus.	Acc. (%)
1	40.97	500	54.86
2	49.31	1000	56.25
4	50.00	2000	56.94
8	49.31	4000	**57.64**
20	52.08	8000	53.47
100	52.78	16000	51.39

Table 2. ADL classification accuracy (Acc, %) using audio-only information, for various GMM classifiers with different numbers of Gaussian mixture components (# Gaus.). The best achieved result, 57.64% audio-only accuracy, is highlighted.

ADL segments with various relative amounts of lost visual tracking. This yielded a so-called "100%" data set, as well as "90%", "50%", "10%", and "0%" sets, the "100%" one containing all ADL segments, the "90%" one only segments whose tracking loss was less than 90% of the total segment duration, and so forth. Details on the size of these sets are given in the last two rows of Table 3.

3.3 Audio-Visual Fusion

In our joint audio-visual ADL classification experiments, we use simple concatenative feature fusion of audio and visual features. This is straightforward to implement, since both audio and visual features are available at 100 Hz (the latter after interpolation). This process yields 16-dimensional joint audio-visual vectors (13 dimensions correspond to PLP features and three to the visual ones).

4 Experiments

Table 2 depicts audio-only ADL classification results using various GMM classifiers trained on PLP features. The best classification accuracy is 57.64%, achieved by a 4000-mixture GMM.

Table 3 depicts visual-only classification results using various GMM classifiers, reported on the various nested data sets discussed above. Note that we always match the training with the test set – e.g. when a GMM classifier is trained on a "10%" reduced training set, it is also tested on a "10%" reduced test set. Here, we observe a clear trend of the "peak" accuracy across the data sets – it moves towards fewer GMM components for the data sets with fewer data. We also observe that the classification accuracy rapidly increases as the data set gets "cleaner". In other words, lost tracking takes a toll on visual-only ADL classification performance.

Finally, Table 4 depicts ADL classification results using the joint audio-visual features on the full data set ("100"). Clearly, the best result (65.97% accuracy) is significantly better than both audio-only (57.64%) and visual-only (46.53%) results.

	Training/Testing Sets				
# Gaus.	"100"	"90"	"50"	"10"	"0"
1	31.94	43.00	44.87	50.98	**65.62**
2	43.75	49.00	48.72	45.10	40.62
4	38.19	**57.00**	47.44	**58.82**	59.38
8	43.06	50.00	**53.85**	52.94	59.38
16	**46.53**	48.00	46.15	**58.82**	56.25
32	45.83	45.00	46.15	54.90	46.88
50	40.97	44.00	39.74	50.98	43.75
100	40.97	44.00	41.03	52.94	43.75
Tr. seg.	576	427	324	226	159
Ts. seg.	144	100	78	51	32

Table 3. Visual-only ADL classification accuracy using various GMMs, trained/tested on a nested sequence of data sets with improving tracking accuracy, moving from the left-most column (full data set) to the right one. The number of segments on the training (Tr.) and test sets (Ts.) are also shown for each condition.

5 Conclusions

Our experiments on the ADL classification show that combining the audio and visual data streams in a multimodal fusion improves the classification accuracy over the ADL classification for each modality separately. Each modality has its own shortcomings: most of the ADLs do not provide acoustic footprint distinctive enough and the data are acoustically sparse making most ADLs seem acoustically similar. Visual data provide much more relevant ADL information, but a larger complexity of the visual data may cause problems extracting the ADL relevant information which is also the case of our work. Moreover the visual data have usually limited coverage in a sense that not all the smart home space is covered, there may be the occlusions etc. This too happens in our work, where there is no visual tracking available in one room. Audio and visual streams are thus complementing each other. We have shown that GMM modeling of the ADL

# Gaus.	Acc. (%)	# Gaus.	Acc. (%)
1	43.75	100	64.58
2	42.36	500	62.50
4	56.25	1000	61.81
8	57.64	2000	61.11
50	**65.97**		

Table 4. ADL classification accuracy, %, using joint audio-visual features for various numbers of GMM components on the full ("100") data set. The best achieved result of 65.97% is highlighted and is clearly significantly better than the audio-only accuracy of 57.64% and visual-only accuracy of 46.53% (see Tables 2 and 3, respectively).

classes is capable to combine information from both streams effectively and improve the classification accuracy.

6 Acknowledgements

The authors wish to thank IBM colleague Larry Sansone for data annotation, as well as Jing Huang (IBM) and Xiaodan Zhuang (UIUC) for work on acoustic scene analysis. In addition, FBK colleagues Massimo Zancanaro, Allesandro Cappelletti, Bruno Lepri, Stefano Messelodi, and Francesco Tobia for the design of the ADL corpus and its collection. We would also like to acknowledge partial support of this work by the European Commission under Integrated Project Netcarity.

References

1. *Netcarity – Ambient Technology to Support Older People at Home.* [Online] Available at: `http://www.netcarity.org`
2. A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," In *Multimodal Technologies for Perception of Humans (CLEAR 2006)*, R. Stiefelhagen and J. Garofolo (Eds.), Springer-Verlag, LNCS 4122, pp. 311–322, 2007.
3. R. Stiefelhagen, K. Bernardin, R. Bowers, R. Travis Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 evaluation," In *Multimodal Technologies for Perception of Humans (CLEAR 2007 and RT 2007)*, R. Stiefelhagen, R. Bowers, and J. Fiscus (Eds.), Springer-Verlag, LNCS 4625, pp. 3–34, 2008.
4. M. Grassi, A. Lombardi, G. Rescio, P. Malcovati, A. Leone, G. Diraco, C. Distante, P. Siciliano, M. Malfatti, L. Gonzo, V. Libal, J. Huang, and G. Potamianos, "A hardware-software framework for high-reliability people fall detection," In *Proc. IEEE Conf. on Sensors*, Lecce, Italy, pp. 1328–1331, 2008.
5. A. Fleury, M. Vacher, H. Glasson, J.-F. Serignat, and N. Noury, "Data fusion in health smart home: Preliminary individual evaluation of two families of sensors," In *Proc. Int. Conf. of the Int. Soc. for Gerontechnology*, Pisa, Italy, 2008.
6. J. Huang, X. Zhuang, V. Libal, and G. Potamianos, "Long-time span acoustic activity analysis from far-field sensors in smart homes," in *Proc. Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Taipei, Taiwan, 2009.
7. C. Wojek, K. Nickel, and R. Stiefelhagen, "Activity recognition and room level tracking in an office environment," in *Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Germany, 2006.
8. A. Cappelletti, B. Lepri, N. Mana, F. Pianesi, and M. Zancanaro, "A multimodal data collection of daily activities in a real instrumented apartment," in *Proc. Works. on Multimodal Corpora: From Models of Natural Interaction to Systems and Applications – Held in Conjunction with the 6th Language Resources and Evaluation Conf. (LREC)*, Marrakech, Morocco, 2008.
9. O. Lanz, P. Chippendale, and R. Brunelli, "An appearance-based particle filter for visual tracking in smart rooms," in *Multimodal Technologies for Perception of Humans (CLEAR 2007 and RT 2007)*, R. Stiefelhagen, R. Bowers, and J. Fiscus (Eds.), Springer-Verlag LNCS 4625, pp. 57–69, 2008.