Modelling the Influence of Cultural Information on Vision-Based Human Home Activity Recognition

Roberto Menicatti, Barbara Bruno and Antonio Sgorbissa
Department of Informatics, Bioengineering, Robotics and System Engineering,
University of Genova, Via Opera Pia 13, 16145 Genova, Italy
(E-mail: roberto.menicatti@dibris.unige.it barbara.bruno@unige.it antonio.sgorbissa@unige.it)

***Abstract*—Daily life activities, such as eating and sleeping, are deeply influenced by a person's culture, hence generating differences in the way a same activity is performed by individuals belonging to different cultures. We argue that taking cultural information into account can improve the performance of systems for the automated recognition of human activities. We propose four different solutions to the problem and present a system which uses a Naive Bayes model to associate cultural information with semantic information extracted from still images. Preliminary experiments with a dataset of images of individuals lying on the floor, sleeping on a futon and sleeping on a bed suggest that: i) solutions explicitly taking cultural information into account are more accurate than culture-unaware solutions; and ii) the proposed system is a promising starting point for the development of culture-aware Human Activity Recognition methods.**

***Keywords*—Human Activity Recognition; Culture-aware Robotics; Ambient Assisted Living**

1. Introduction

All human beings are cultural beings. *Culture* is defined as the shared way of life of a group of people, that includes beliefs, values, ideas, language, communication, norms and visibly expressed forms such as customs, art, music, clothing, food, and etiquette. Culture influences individuals' lifestyles, personal identity and their relationship with others both within and outside their culture [1]. Building on these premises, a recent research trend explores the influence of people's culture on their relationship with robots, aiming at assessing its impact on factors, such as acceptability and trust, which are of crucial importance for all applications of robots as personal assistants [2], [3], [4], [5].

In the context of Ambient Assisted Living (AAL), the recognition of human daily activities (and, in particular, of the Activities of Daily Living identified by gerontologists as tightly correlated with a person's autonomy [6], [7]) is crucial to assess the health status of the assisted person. To this aim, vision-based Human Activity Recognition (HAR) systems are gaining more and more importance. Normally, HAR is performed on video streams rather than still images, as shown in some detailed surveys [8], [9]. However, methods based on video streams usually only consider the human silhouette (through tracking or background subtraction) or are based on local small regions of interest (corresponding to different motion patterns) [10], thus disregarding the additional relevant semantic information which can be deduced by the environment surrounding the person performing the action.

It is an established fact that Activities of Daily Living, e.g. *eating* and *sleeping*, are carried out in different ways and different places of the house, in accordance with the cultural identity of the person [11]. Mulholland and Wyss [12] report that, for example, in many parts of Asia, postures such as squatting, kneeling or sitting cross-legged on the floor are more common than using a chair. In Japan a kneeling posture is commonly adopted to perform daily activities such as eating, socializing, and religious or traditional ceremonies such as the tea ceremony. In Asia and the Middle East people sit cross-legged on mats and tatami for resting, socializing, eating, working, or leisure or spiritual activities such as yoga.

Conversely, postures assuming a direct contact with the floor are generally uncommon in European countries and are often associated with potentially dangerous situations (such as a sudden illness, fall, or faint).

Since culture influences and pervades most of the actions of a person, and particularly everyday activities such as eating, sleeping and toileting, we argue that in-home assistive robots should not only be *culture-aware* when directly interacting with a person, but also, and more generally, able of evaluating any type of user-related information in light of said person's culture and preferences.

As a preliminary step towards the development of culture-aware HAR systems, we address the problem of: i) determining whether a person is sleeping or lying in a potentially dangerous situation; ii) taking into account the influence of culture on the way in which the *sleeping* activity is performed, in particular by considering the case in which the person sleeps on a bed, as it is common in European countries, and the case in which the person sleeps on a futon, as it is common in Japan. More precisely, the contribution of the article is two-fold: i) the enhancing of the Cloud-based HAR framework presented in [13] to include cultural information and increase the chances of a right classification; and ii) the comparison of the results obtained by taking culture into account at different levels for the (vision-based) recognition of a person who is sleeping or lying in a potentially dangerous situation.

The article is organized as follows. Section 2 outlines the problem statement and proposes four different solutions for embedding cultural information in the process of recognizing daily activities. Section 3 describes the method we propose for including cultural information both in the training and in the testing phase of a vision-based HAR system. Section 4

compares the tests performed and the results obtained by using the different solutions adopted. Conclusions follow.

2. Problem Statement

Human Activity Recognition systems based on visual information usually require the execution of two distinct phases: during the *training* phase, a number of examples (i.e., labelled images) of the activity to recognize are used for the creation of its model; then, during the *testing* phase, an unlabelled recording (i.e., a new image) is analysed in light of the available models, and labelled as an instance of the one that better matches it. In this context, it is possible to envision different solutions for modelling and recognizing activities in which cultural factors play a non-negligible role.

1) *Individual-specific*. Trivially, if all examples used in the training phase, as well as all the recordings used in the testing phase, belong to one and the same person, the system will always be aligned with that person's culture. This solution, which best captures the unique cultural traits of an individual, requires a long set up and does not allow for exploiting similarities among different persons.
2) *Culture-unaware*. At the opposite end of the spectrum with respect to the individual-specific solution, culture-agnostic systems rely on a large number of examples, from many different individuals, for the creation of models general enough to be valid for different cultures. For example, in the case of the *sleeping* activity, mixing examples from Japanese and European individuals in the training phase might lead to the creation of a model which does not rely on the presence of a bed. This solution minimizes the set up time, since the training is done only once, but, arguably, at the expenses of a reduced accuracy in capturing person-specific traits.
3) *Culture-aware training*. A more interesting solution envisions the creation of culture-specific models of all activities in which cultural factors may play a non-negligible role, thus leading, for example, to the creation of two models of the *sleeping* activity, e.g., *sleeping-futon* and *sleeping-bed*, respectively for the Japanese and European culture. In the testing phase, recordings of a European person sleeping are likely to better match the *sleeping-bed* model, while recordings of a Japanese person sleeping are likely to be more often labelled as occurrences of the *sleeping-futon* activity. This solution builds on the assumption that it is possible to achieve a good trade-off between recognition accuracy and generality of the models by taking cultural differences into account explicitly during the training phase and implicitly in the testing phase, when the system relies on sensory cues only to infer the culture of the person.
4) *Culture-aware training and testing*. Let us consider the image of a person lying on the floor. In the absence of explicit information about his/her cultural profile, the HAR system might lack clear evidence to discriminate between a Japanese person sleeping on a thin futon and a European person who fell and is in need of assistance. This solution builds on the assumption that, by explicitly considering cultural information during both the training and the testing phase, it is possible to improve the system accuracy, with no loss in the generality of the models.

The first solution, which does not specifically address the problem of modelling the influence of culture on daily activities, is not considered in this article. The interested reader might find relevant information about individual-specific HAR systems in [13].

The second and third solutions allow for the adoption of any vision-based HAR method, since the cultural information is encoded in the images collected for the training phase, and, in the third solution, explicitly expressed by creating different, culture-dependent, models of the same activity.

Conversely, to the best of our knowledge, there is no vision-based HAR system allowing for explicitly considering cultural information during the testing phase. The following Section outlines the method we propose to this aim, together with the rationale for the design choices supporting it.

3. Method

Albeit scarce, it is possible to find in the literature examples of robotic systems which model cultural information to tune their behaviour towards an individual. Torta et al. propose a method to parametrize the interpersonal distance and direction of approach that the robot should use when talking to a person [14]. Information about the acceptability of different values of distance and orientation is encoded in a multi-dimensional function and combined with contingent sensory information (for example, concerning the presence of obstacles) in a Bayesian inference mechanism with a particle filter to identify a suitable target pose for the robot.

A more complex example describes a framework for the learning and selection of culturally appropriate greeting gestures and words [15]. In the proposed system, an initial set of gestures and words is extracted from video and text corpora, and initial associations between gestures and words and a number of cultural factors of relevance are taken from literature in social studies and expressed as conditional probabilities in a Naive Bayes classifier. At run-time, the user's cultural profile in terms of the cultural factors is computed and used to identify the greeting gestures and words which better match the profile.

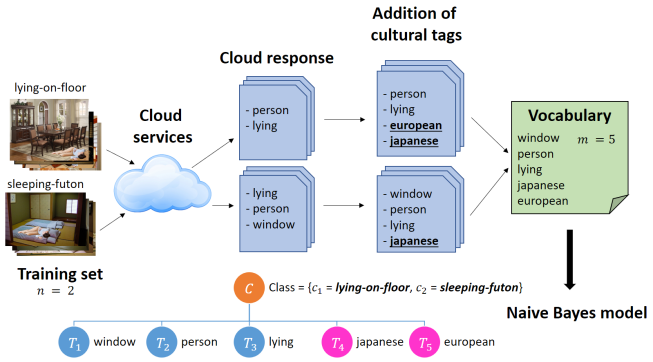
In accordance with literature findings, we propose the use of a Naive Bayes classifier for associating cultural information to visual features.

In particular, in this work we rely on the Cloud-based HAR (CHAR) framework presented in [13] and extend it to include cultural information. CHAR exploits the computer vision cloud services provided by Clarifai[1], Microsoft[2] and Google[3] to extract semantic information from static images. In particular, the three cloud services return a list of tags describing the objects, the environments, the actions etc.

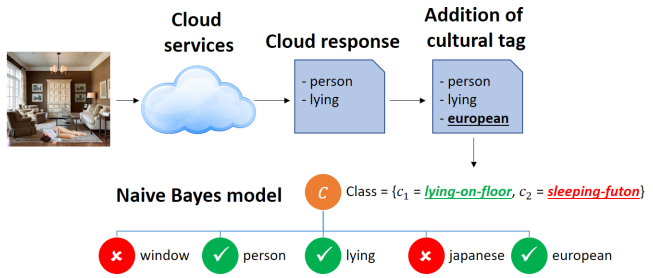
[1] http://www.clarifai.com/

[2] https://azure.microsoft.com/it-it/services/cognitive-services/computer-vision/

[3] https://cloud.google.com/vision/



(a) Training



(b) Testing

Fig. 1. Architecture of the training/testing phase of the proposed HAR system with culture aware training and testing

detected in the image without giving any information about their spatial relation. As Fig. 1(a) shows, during the training phase a number of training images for each activity of interest are available. The tags returned by the three cloud services for all the training images are collected and used to train a Naive Bayes model, in which the parent node represents the activity (henceforth also referred to as the *class*) and each child node is a tag extracted from the training sets. An activity is therefore modelled as a probability distribution over the whole set of possible tags. During the testing phase (see Fig. 1(b)), an image is given to the cloud services to retrieve the associated tags, which are then compared with the available models to identify the activity more likely to be represented in the image.

A number of reasons support the choice of this method as a starting point for the development of a HAR framework with culture-aware training and testing: i) by relying on a Naive Bayes model and semantic tags, it is very close to the aforementioned methods proposed in the literature for the modelling of cultural factors; ii) it can be easily adapted to match the requirements of the *culture-unaware* and *culture-aware training* solutions, to allow for a meaningful comparison of the different approaches; iii) it is based on publicly accessible online services, thus enforcing reproducibility. Figure 1(a) on the right side shows our proposal for enhancing the CHAR system with cultural information. In particular, in the training phase to the list of tags returned by the cloud services for each training image, we add a tag describing the cultural identity of the person depicted in the image, assuming that this information is available for all the images in the training set. Let us refer to this kind of tag as *cultural tag*. The vocabulary will then include the cultural tags, thus extending the Naive Bayes model with additional children *cultural nodes* (shown in fuchsia in Fig. 1(a)), one for each cultural tag included.

As for the other tags, the cultural tags have a different probability distribution for each trained class. If a class represents a culture-specific activity or situation (e.g. sleeping on a futon), the probability value of the corresponding cultural tag (e.g. a tag *japanese*) for that class is equal to 1 while all the other cultural tags, representing other cultures, (e.g. tags as *italian*, *mexican*...) have probability 0. If, instead, the class is culture-independent because the scene represented is an activity which is performed in the same way in different countries the probability distribution of the cultural tags will be proportional to the measure of the presence of each culture considered in the training class. Consider, for example, a class *reading* which contains 6 images of different persons reading. In 3 images the person is Italian, in 1 is Japanese and in 2 images is Mexican. The probability of the corresponding cultural tags *italian*, *japanese* and *mexican* for that class are $3/6$, $1/6$ and $2/6$ respectively.

In the testing phase (Fig. 1(b)) a tag with the cultural profile of the person shown in the tested image is added to the list of tags returned by the cloud services. Also in this case it is safe to assume that such cultural knowledge is available, since a device performing image recognition in an AAL scenario (e.g. an assistive robot) can be given the knowledge of the cultural identity of the user during its setup. The presence of a cultural tag and the absence of the other ones will set the states of the relative cultural nodes, thus influencing the classification of the image.

4. Tests and Results

As anticipated in Section 2, in this work we focus only on the second, third and fourth solutions described, that we denote as:

- *Culture-Unaware* (CU);
- *Culture-Aware Training* (CAT);
- *Culture-Aware Training and Testing* (CATT).

To address the problem of determining whether a person is sleeping or lying in a potentially dangerous situation, taking cultural information into account, we consider the following three situations:

1) sleeping on a bed (associated to European culture);
2) sleeping on a futon (associated to Japanese culture);
3) lying on the floor (not associated to any specific culture).

Discriminating one of such situations from the others is not trivial: all such activities involve a person in the same posture (i.e., lying) and two of them ("sleeping on a futon" and "lying on the floor") are strikingly similar. A wrong classification could, for example, classify a person who is sleeping on a futon as someone who has fallen and, in the context of AAL, result in a false alarm raised by the robot, or

monitoring system, and thus a reduced reliability. The opposite misclassification has even worse consequences: a robot might fail to alert of a fall, by wrongly classifying a person lying on the floor as one sleeping on a futon.

All of our tests have been performed offline; in particular, we have collected a dataset of 36 images (Fig. 2), divided in:

- class *sleeping*, 24 images divided in:
  - subclass *sleeping-bed*, 12 images;
  - subclass *sleeping-futon*, 12 images;
- class *lying-on-floor*, 12 images.

To further increase the ambiguity, all the images were obtained by putting the same figure of a lying-down person over different backgrounds, thus visually simulating the three different cases. While the images of subclasses *sleeping-bed* and *sleeping-futon* have respectively European and Japanese home backgrounds only, for the *lying-on-floor* class we have collected 6 images with a European home background and 6 images with a Japanese home background.

4.1. Culture-Unaware

As defined in Section 2, a culture-unaware solution does not take into explicit consideration the different ways in which a same activity is performed according to different cultures. Therefore, in this case we consider a training dataset composed of 12 images for each class (i.e., "*lying-on-floor*" and "*sleeping*"). The training set of the *sleeping* class is equally split among its two subclasses.

We have adopted the standard CHAR system for the recognition, and used *k-fold cross validation* upon the dataset for the training and testing. We have randomly divided each class in 3 subsets of 4 images each, then, by taking one subset of each class for the testing set and the remaining two subsets of each class for the training set, we have combined them in 9 possible combinations. Therefore, each subset is used 3 times for the testing and 6 times for the training.

4.2. Culture-Aware Training

The CAT solution assumes an explicit modelling of cultural information in the training phase only. In this case we use the full dataset, divided in the three classes mentioned before. During the training, the cultural information is taken into account by considering the *sleeping* class as made of the two, independent subclasses *sleeping-bed* and *sleeping-futon* which allows for the separation of the different ways the act of sleeping is performed and the different ways this situation is depicted in the images.

As for the CU solution, also in this case we have used the standard CHAR system and k-fold cross validation. The 3 subsets of each class have been combined in 27 folds using, at each fold, one subset for testing and two for training for each class. Each subset is used 9 times for the testing and 18 times for the training. All subsets of the *lying-on-floor* class contain two images with a "European" background and two images with a "Japanese" background.

TABLE 1
PROBABILITY DISTRIBUTION OF THE ADDITIONAL CULTURAL TAGS *european* AND *japanese*

	european		*japanese*	
	present	absent	present	absent
sleeping-bed	1	0	0	1
sleeping-futon	0	1	1	0
lying-on-floor	0.5	0.5	0.5	0.5

4.3. Culture-Aware Training and Testing

The CATT solution assumes an explicit modelling of cultural information not only in the training phase but also at the moment of classifying an image. We have again used k-fold cross validation, with the same folds of the CAT solution. However, instead of using CHAR, we have used its modified version, described in Section 3. The modified version of CHAR has been implemented for the training and testing of this specific case as follows.

A. Training: The training set of each class consists of 8 images. In accordance with the method explained in Section 3, we have added a cultural tag to all the training images. In particular, we have added the tag:

- *european* to all the images of the class *sleeping-bed* and to the images of the class *lying-on-floor* with the "European" background;
- *japanese* to all the images of the class *sleeping-futon* and to the images of the class *lying-on-floor* with the "Japanese" background.

The probability distribution of the two cultural tags is shown in Table 1. Please notice that the values in the table are representative of how we have built the training set for the purpose of assessing the performance of the solution and do not represent any real probability distribution of these cultural factors.

The two cultural tags and their probabilities are automatically included in the training of the Naive Bayes model.

B. Testing: The testing set of a class consists of 4 images. The two cultural tags have been added in the same way as for the training.

The size of the training set (8 pictures) and testing set (4 pictures) of a class is the same for all the methods. This, of course, leads to a different overall number of test samples for the three solutions when using k-fold cross validation, since the number of classes is 2 in the CU case and 3 in the CAT and CATT ones. However, it allows for a better comparison of the performance of the three methods over each single class of image.

4.4. Results

In order to evaluate the performance of the three different solutions we have computed their confusion matrices, shown in Fig. 3. With reference to the figure, the columns correspond to the actual class of the tested images (target class) and the rows represent the predicted class (output class). The diagonal

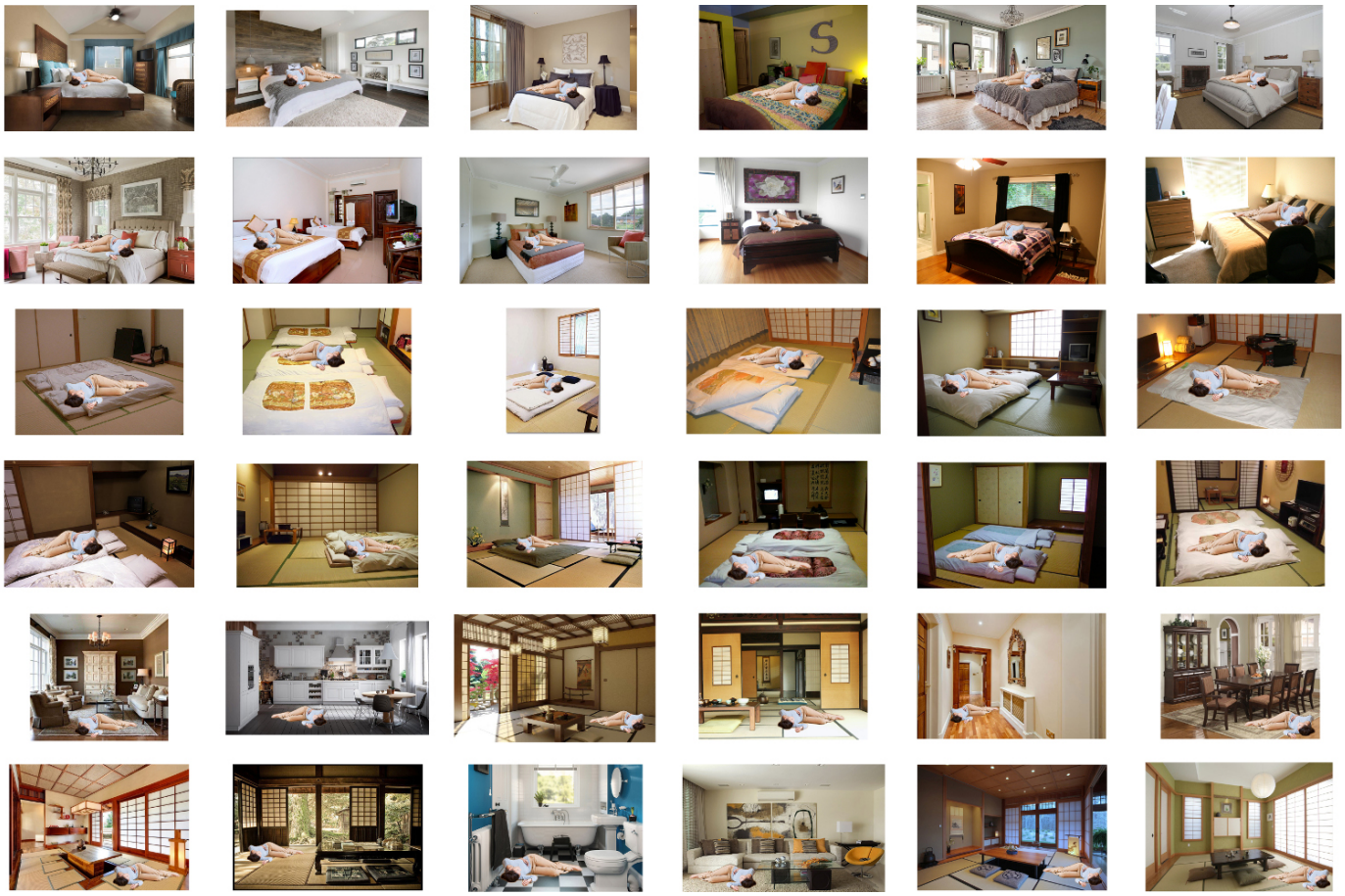
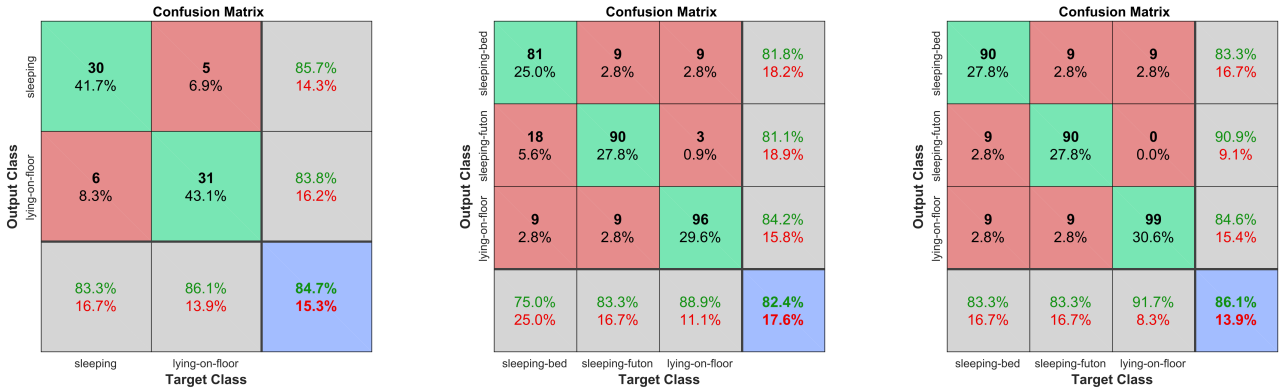


Fig. 2. Dataset: subclass *sleeping-bed* (rows 1-2), subclass *sleeping-futon* (rows 3-4), class *lying-on-floor* (rows 5-6)



(a) CU (b) CAT (c) CATT

Fig. 3. Confusion matrices for CU (a), CAT (b), CATT(c) solutions

cells (green background) show the number of True Positives of each class, i.e. the number of images which have been classified correctly, and the percentage over the overall number of images in all the testing sets. The off diagonal cells show, instead, the number of wrong detections. The last row shows the *recall* (or true positive rate) of each class, the last column shows the *precision* of each class, the bottom-right cell (blue background) shows the overall accuracy.

By comparing the two confusion matrices for solutions CAT and CATT we can see that 9 *sleeping-bed* images which are wrongly classified as *sleeping-futon* in the case without cultural tags are, instead classified correctly when these tags are included (cells of first and second row, first column of Fig. 3(b) and Fig. 3(c)). Moreover, the 3 *lying-on-floor* images wrongly classified as *sleeping-futon* in the first case, are then classified correctly when including the cultural tags (cells of second and third row, third column of Fig. 3(b) and Fig. 3(c)).

Moreover, for those images which both CAT and CATT solutions have classified incorrectly, we have computed the difference in the confidence scores and averaged over the different images. We have noticed that the CATT solution presents an average confidence score for the misclassified images lower by 5% than the one of the CAT solution.

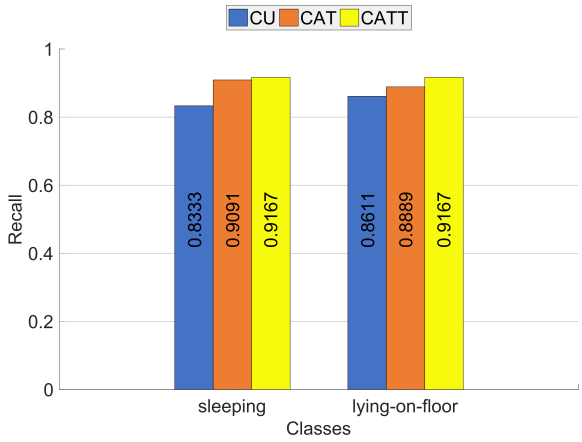
In order to better compare the performance of CAT and CATT with CU we have grouped the results of the subclasses *sleeping-bed* and *sleeping-futon* for the case CAT and CATT as if they were a single class *sleeping* and computed the aggregate recall and precision values. With reference to the confusion matrices of Fig. 3(b) and Fig. 3(c), let us call M indifferently one of the two matrices, i the row index and j the column index. Concretely, for the superclass *sleeping* both in the case of CAT and CATT we have computed:

$$recall_s = \frac{TP_s}{TP_s + FN_s} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} M_{ij}}{\sum_{i=1}^{2}\sum_{j=1}^{2} M_{ij} + \sum_{j=1}^{2} M_{3j}} \quad (1)$$

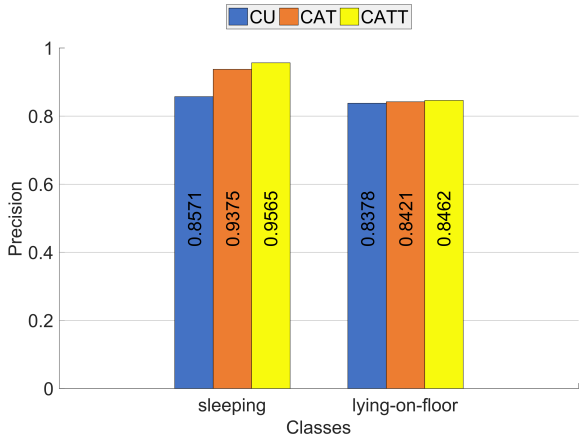
$$precision_s = \frac{TP_s}{TP_s + FP_s} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} M_{ij}}{\sum_{i=1}^{2}\sum_{j=1}^{2} M_{ij} + \sum_{i=1}^{2} M_{i3}} \quad (2)$$

where *TP* stands for *true positives*, *FN* stands for *false negatives*, *FP* stands for *false positives* and the subscript *s* stands for the class *sleeping*.

Finally, the plots in Fig. 4 compare, for each class, the recall and the precision obtained through the three different solutions adopted. As expected, the figures show that adding cultural information improves the recognition performance, both in terms of precision and recall, with the CATT solution being the most accurate.



(a) Recall



(b) Precision

Fig. 4. Comparison of recall (a) and precision (b) between the three solutions (CU, CAT, CATT)

5. Conclusions

In this article, we have discussed the influence of a person's culture on his daily activities, and, as a consequence, the need for taking cultural information into account when designing automated Human Activity Recognition systems.

In the field of vision-based HAR systems for the recognition of Activities of Daily Living, we have identified three possible non-trivial solutions to the problem, and compared their performance. Of these three solutions, one attempts to create general models by involving examples from people belonging to different cultures in the same training class (CU), while the other two explicitly include cultural clues either in the training (CAT) or both in the training and testing phase (CATT).

As a possible implementation of the last approach, we have adopted a system for the recognition of daily activities and indoor scenes which relies on cloud-based computer vision services and a Naive Bayes model to represent activities as distributions of probabilities over a set of tags, and enhanced it with tags encoding relevant cultural information.

To compare the performance of the CU, CAT and CATT solutions, we have collected a dataset of images pertaining to two activities (i.e., sleeping and lying on the floor as a consequence of a fall, or a sudden illness), and distinguished between Japanese people sleeping on a futon, and European people sleeping on a bed.

Experiments support our hypothesis that: i) explicitly taking cultural information into account allows for a more accurate recognition; ii) the proposed method is a suitable choice as a culture-aware HAR system, with an overall precision above 84% and an overall recall above 91%.

6. Acknowledgement

This work has been partly supported by the European Commission Horizon2020 Research and Innovation Programme under grant agreement No. 737858 (CARESSES).

References

[1] I. Papadopoulos, *Transcultural health and social care: development of culturally competent practitioners*. Elsevier Health Sciences, 2006.

[2] G. Trovato, M. Zecca, S. Sessa, L. Jamone, J. Ham, K. Hashimoto, and A. Takanishi, "Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by egyptians and japanese," *Paladyn, Journal of Behavioral Robotics*, vol. 4, no. 2, pp. 83–93, 2013.

[3] G. Eresha, M. Häring, B. Endrass, E. André, and M. Obaid, "Investigating the influence of culture on proxemic behaviors for humanoid robots," in *2013 IEEE RO-MAN*. IEEE, 2013, pp. 430–435.

[4] M. P. Joosse, R. W. Poppe, M. Lohse, and V. Evers, "Cultural differences in how an engagement-seeking robot should approach a group of people," in *Proceedings of the 5th ACM international conference on Collaboration across boundaries: culture, distance & technology*. ACM, 2014, pp. 121–130.

[5] S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr, "Effects of culture on the credibility of robot speech: A comparison between english and arabic," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 157–164.

[6] S. Katz, A. Chinn, and L. Cordrey, "Multidisciplinary studies of illness in aged persons: a new classification of functional status in activities of daily living," *Journal of Chronic Disease*, vol. 9, no. 1, pp. 55–62, 1959.

[7] M. Lawton and E. Brody, "Assessment of older people: self-maintaining and instrumental activities of daily living," *The Gerontologist*, vol. 9, pp. 179–186, 1969.

[8] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2013.01.013

[9] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human Action Recognition With Video Data: Research and Evaluation Challenges," *Ieee Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

[10] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[11] M.-W. Zeenat, "Why this interest in minority ethnic groups?" *British Journal of Occupational Therapy*, vol. 59, no. 10, pp. 485–489, 1996. [Online]. Available: http://dx.doi.org/10.1177/030802269605901009

[12] S. J. Mulholland and U. P. Wyss, "Activities of daily living in non-western cultures: range of motion requirements for hip and knee joint implants," *International Journal of Rehabilitation Research*, vol. 24, no. 3, pp. 191–198, 2001.

[13] R. Menicatti and A. Sgorbissa, "A cloud-based scene recognition framework for in-home assistive robots," submitted to 2017 IEEE International Symposium on Robot and Human Interactive Communication.

[14] E. Torta, R. H. Cuijpers, J. F. Juola, and D. van der Pol, "Design of robust robotic proxemic behaviour," in *International Conference on Social Robotics*. Springer, 2011, pp. 21–30.

[15] G. Trovato, M. Zecca, M. Do, Ö. Terlemez, M. Kuramochi, A. Waibel, T. Asfour, and A. Takanishi, "A novel greeting selection system for a culture-adaptive humanoid robot," *International Journal of Advanced Robotic Systems*, vol. 12, 2015.