Bridging Vision and commonsense for Multimodal Situation Recognition in Pervasive Systems

Nicola Bicocchi
Dip. di Ingegneria dell'Informazione
Universitá di Modena e Reggio Emilia
Modena, Italia
nicola.bicocchi@unimore.it

Matteo Lasagni
Institute of Computer Engineering
University of Lübeck
Lübeck, Germany
lasagni@iti.uni-luebeck.de

Franco Zambonelli
Dip. di Scienze e Metodi dell'Ingegneria
Universitá di Modena e Reggio Emilia
Reggio Emilia, Italia
franco.zambonelli@unimore.it

***Abstract*—Pervasive services may have to rely on multimodal classification to implement situation-recognition. However, the effectiveness of current multimodal classifiers is often not satisfactory. In this paper, we describe a novel approach to multimodal classification based on integrating a vision sensor with a commonsense knowledge base. Specifically, our approach is based on extracting the individual objects perceived by a camera and classifying them individually with non-parametric algorithms; then, using a commonsense knowledge base, classifying the overall scene with high effectiveness. Such classification results can then be fused together with other sensors, again on a commonsense basis, for both improving classification accuracy and dealing with missing labels. Experimental results are presented to assess, under different configurations, the effectiveness of our vision sensor and its integration with other kinds of sensors, proving that the approach is effective and able to correctly recognize a number of situations in open-ended environments.**

***Keywords*-pervasive computing; situation recognition; commonsense knowledge; image analysis; activity recognition; mobility.**

I. Introduction

Situation-awareness, i.e., the capability of automatically recognizing the situations in which an entity is engaged, is a key feature to be enforced in pervasive computing services, to enable them to dynamically adapt their behavior to the current situation. Beside location-based mobile applications and services, which must exhibit forms of situation-awareness, the need to acquire situation-awareness arises in a variety of other scenarios like mobile robots, smart buildings, e-vehicles and adaptive traffic management.

Due to its increasing importance, many researchers in pervasive computing are working on situation recognition. Most of them focus on algorithms and tools to identify specific situational aspects (e.g., location [18], current activity [14], health status [4], daily routines [31], [9] etc.) from sensor data. In particular, the key goal is to analyze and classify the data produced by available sensors. This can take place by analyzing and classifying the data of individual sensors or, when multiple sensors are available, in a multimodal way [30] [21], by fusing together the classification results of different sensors to increase accuracy in recognition, enabling complex multi-faceted situation recognition.

Among many works attacking situation recognition using sensors such as accelerometers, magnetometers, GPS, microphones, light sensors, and so forth, vision (i.e., cameras as sensors) has not received the same degree of attention so far. On the one hand, vision is potentially very informative: just think of how much information human brain is capable of extracting from what we see. On the other hand, low-cost miniaturized cameras have already hit the market, and we can reasonably expect that micro cameras continuously capturing what is around will be part of our future everyday life-logging toolset.

In our opinion, one of the most relevant obstacle that prevented vision to be intensively exploited as a situation-recognition tool in pervasive computing so far is the lack of reliable and general-purpose classifiers. In fact, while it is simple to acquire images, it is still cumbersome to detect and concurrently classify a wide range of objects, people, and scenes. Many attempts have been made in the recent past [10]. These latter focused on recognizing specific classes of objects through complex parametric representations, preventing vision to be an attractive source of information in pervasive systems due to their computational requirements.

Recent advances in image processing, however, have shown that it is possible to implement effective general-purpose image classifiers using non-parametric algorithms [26], [22], [15]. With a large datasets of loosely classified images, it is possible to find, with high probability, images visually similar to a query image, containing alike scenes with related objects arranged likewise. Then, even though the images in the retrieval set are only partially labeled (as tagged Web images typically are), it is still possible to propagate the labels to the query image, so as to perform classification.

Starting from this idea, which has proved to be effective under controlled conditions, we have extended and integrated it into a more elaborate framework exploiting commonsense knowledge, suitable for multimodal situation recognition in pervasive computing scenarios.

In particular, the contributions of this paper are as follows:

- We detail the design and implementation of our vision sensor. It is based on deconstructing an image to individual entities/objects, classify them with a non-parametric approach and then on reconstructing a comprehensive classification out these individual features with commonsense knowledge. Experiments show that our vision sensor can represent an effective tools to extract compact yet expressive situational information (i.e., the scene) in an unattended way from a simple camera.
- We show how, using a simple framework based again on commonsense knowledge, the vision sensor can be effectively integrated with other sensors for multimodal situation recognition, so as to improve classification accuracy and deal with missing labels.

The remainder of this paper is organized as follows: Section II describes how the vision sensor has been designed and developed. Section III assesses the effectiveness of the vision sensor in different types of situations, and discusses its current limitations. Section IV details our commonsense approach for multimodal sensor fusion, and quantifies its benefits with experiments. Section V discusses related work. Section VI concludes the paper and outlines future developments.

II. The Vision Sensor

A miniaturized wearable camera can be easily embedded in a suite and configured to automatically collect user's environment pictures. Because of this, they could potentially already be used as sensors enabling situation recognition.

Results presented in [26] and [22] showing that it is possible to recognize objects, persons and scenes using large datasets combined with non-parametric algorithms. Such results inspired us to investigate how to integrate vision in pervasive systems. However, we immediately faced a notable problem: images sampled in unpredictable conditions might be confused or not very informative. The camera might be out of focus, too close to the target, partially hidden by an obstacle, pointing to a dark spot, and so forth.

These premises led us to develop a technique able to make use of a number of low-quality, and eventually low-meaning images to perform situation and scene recognition. Specifically, instead of performing classification on a single good-quality image, we classify a set of images (i.e., the temporal window in which images have been captured). The process is based on the extraction of a number of meaningful details from a set of images, their independent classification, and the recognition (i.e., classification) of the global scene using commonsense knowledge. The overall process is depicted in Figure 1 and can be summarized as follows:

0) *Image sampling*: images are sampled at specific rates to obtain different samples of the overall surrounding scene.
1) *Unsupervised image segmentation*: each image is segmented using an unsupervised technique to extract individual features (e.g. objects) form the overall image, thus fragmenting the overall classification problem into the problem of classifying a set of simpler sub-images uniformly scaled to a tiny size.
2) *Image classification*: each sub-image is classified, accordingly to [26], by searching its nearest neighbors within a loosely labelled dataset of $7.9x10^7$ images (a). For each sub-image, and based on the labels of its nearest-neighbors, a voting scheme on the Wordnet knowledge base is applied to associate a label to it (b).
3) *Situation classification*: labels assigned to images (and segmented sub-images) taken within a temporal window are positioned within the ConceptNet knowledge base as *label nodes*. Situation classification is performed by searching the *class node* with the shortest average path to the individual *label nodes* (see Figure 2) so as to eventually identify a label that re-construct a global comprehensive situation (i.e., a scene).

As a result, a simple camera can become an effective and usable vision sensor, based on which users can compactly classifying situations around. For example, as in Figure 1, by recognizing that the user is currently in a living room.

Let us now detail each of the above phases.

A. Unsupervised Image Segmentation

Image segmentation (Figure 1, step 1), is a key stage in our approach because of two motivations. The former is that randomly-taken images usually have low quality and semantic levels. Because of this, searching them within a dataset, even within 79 million images, can be unfruitful. The latter, instead, is related to the very low resolution of the images of the dataset we have used (i.e. 32x32 pixels). Moreover, in order to search for neighbors within this dataset, query images have to match this size. Despite this resolution is sufficient to perform classification of a scene, a person or an object [17] it is only possible if the majority of visible area actually contains the subject. In our case, randomly-taken images, might contain informative details that would be completely lost by shrinking image size to 32x32 pixels. Due to these considerations, we implemented a modified version of Grabcut to isolate the key components of each image.

Grabcut [23] is an efficient tool for background suppression. Removing the background surrounding each element simplifies the contours elaboration of each separable subject and, as consequence, its bounding rectangle. To discern automatically between background and foreground of a single image, it is necessary to define a model of at least one of them. Grabcut has been designed as an interactive tool requiring users to select some areas forming the background

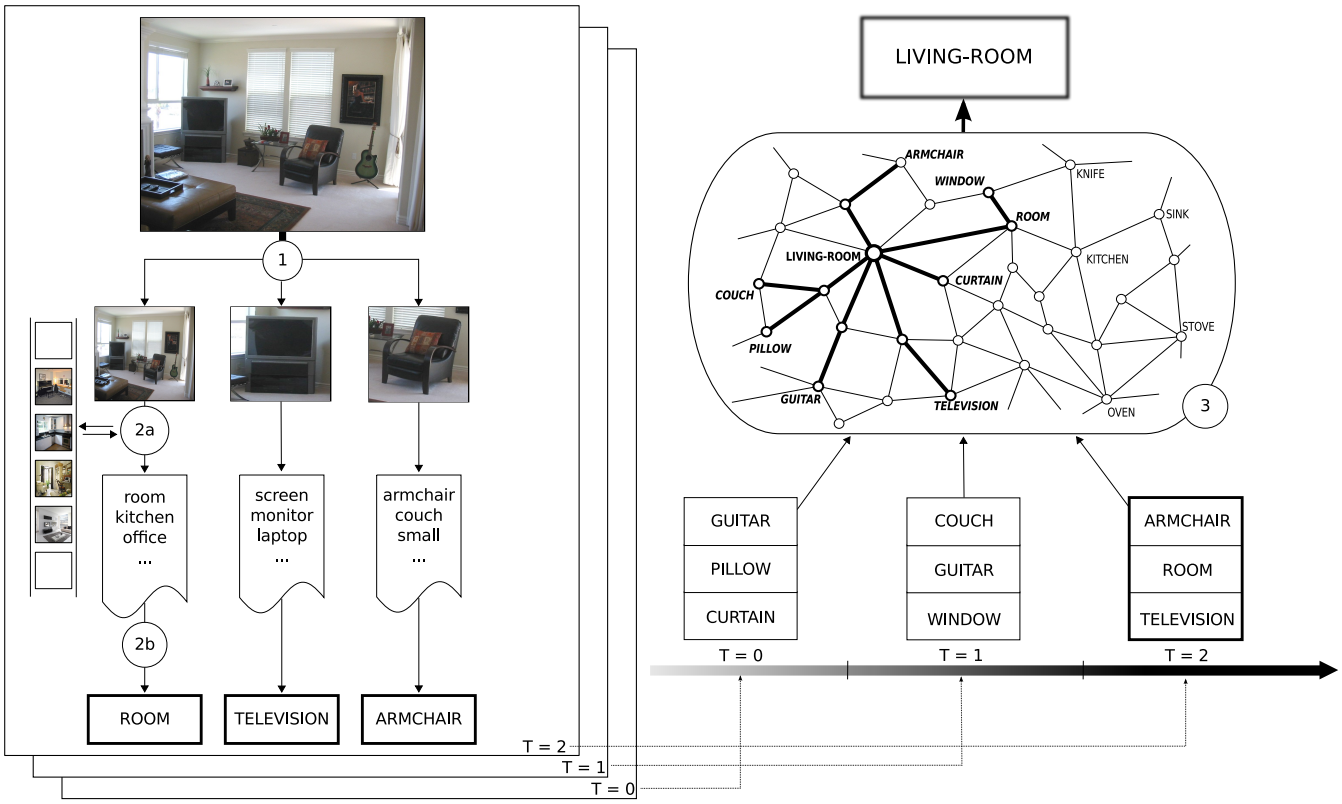

Figure 1. The main phases towards classification in our vision sensor. *(1) Image segmentation*: images periodically taken by the camera are segmented using an unsupervised technique to extract individual objects/entities from the overall scene; *(2) Image classification*: each sub-image is classified by searching its nearest neighbors within a loosely labelled large dataset of images (a) and applying a voting scheme on the Wordnet knowledge base (b); *(3) Situation classification*: labels assigned to images (and to segmented sub-images) from within a given temporal sliding window are positioned within the ConceptNet knowledge base as *label nodes*. Situation classification is performed by searching the *class node* with the shortest average path to the *label nodes*, so as to obtain a re-constructed comprehensive classification of the overall scene.

model. To enable this process to run without supervision, we came to the assumption that generally the background surrounds the subjects that are likely to lay in the center of the frame. Therefore, we consider an outline 5% inwards respect to the border of the image, that delimits the initial pixels belonging the background. Once the background model is determined, Grabcut retrieves all the remaining areas composing the background itself. We are conscious that assuming that the background model is always situated at the border of the frame produces segmentation errors. Nevertheless, if the background model contains an object, the only consequence is that that object will not be segmented and, thus, ignored in the following phases. Considering that our approach is based on aggregating and processing a number of different views of the same scene, this results acceptable.

Given one input image, this stage produces a set of 32x32 images composed by the initial image itself and all sub-images that have been identified by the segmentation algorithm.

B. Image Classification

Each tiny image is then classified by applying the process detailed in [26]. Here we provide a small summary. We use a visual database containing $7.9x10^7$ images (32x32, 24-bit). Each image is loosely labeled with a keyword belonging to the Wordnet knowledge base. Due its size, the dataset can not be manually labelled and optimized. Thus, it contains noise in terms of duplicate images and not relevant labels. Furthermore, duplicate images might have different labels. The classification process can be decomposed in two parts: neighbors search and voting.

Neighbors are searched (see Figure 1, step 2a) by executing an exhaustive search on the whole dataset. Images are represented as ordered vectors of 3072 elements, normalized to have zero mean and unit norm. Then, the sum of squared differences is computed:

$$\sum_{x,y,c}(I_1(x,y,c) - I_2(x,y,c))^2$$

This process is extremely expensive in terms of computational resources. To begin, we made all the computations offline. Nevertheless, recent in-memory indexing techniques

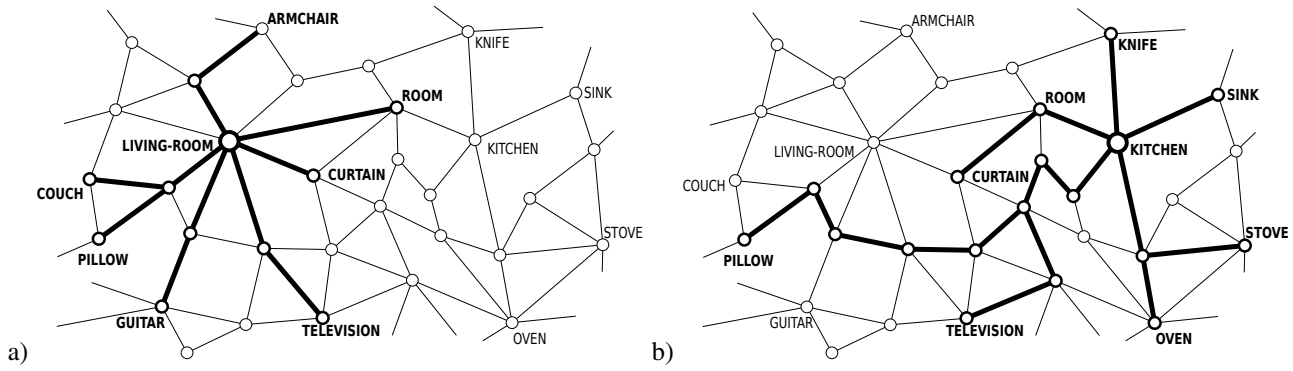

Figure 2. Situation classification within ConceptNet. Each class candidate for classification is marked on ConceptNet as a *class node* C_i (living room, kitchen); each label coming from the image classification phase is marked as a *label node* l_j. For each *class node* C_i is computed the average shortest path pl_i to every *label node* l_j. Classification is performed by selecting the class C_i with the lowest pl_i. In this figure the transition process between two domestic environments is exemplified.

such as spectral hashing [29] can be used to speed up the process of several magnitude orders.

Given a set of neighbors, each one associated to a label, the classification is performed by voting on the Wordnet tree (see Figure 1, step 2b). Wordnet provides semantic relationships between the 75,062 nouns for which we have collected images. For simplicity, we reduce the initial graph-structured relationships between words to a tree-structured one by taking the most common meaning of each word. The result is a large semantic tree whose nodes consist of the 75,062 nouns and their hypernyms, with all the leaves being nouns.

Given a query image, each neighbor image votes for its branch within the tree. In this manner votes are accumulated across a range of semantic levels and the effects of the labeling noise are averaged out over many neighbors. Classification is performed by assigning to the query image the label with the most votes at the desired height (i.e. semantic level) within the tree. The number of votes acts as a measure of confidence.

The result of this process is a set of keywords, each one associated to an image, or sub-image, taken in input.

C. Situation Classification

Classifying a single image is not enough for our purposes. As already mentioned, images sampled in unpredictable conditions are often not informative. Additionally, the segmentation phase described above actually splits the global situation in a number of details. For example, a living room can be associated to a set of labels such as *table, chair, lamp, mirror, window, television*.

To mitigate this problem, we considered a set of images collected within a temporal window instead of a single one. We can safely assume that every user spends a minimum amount of time (i.e. minutes at least) in a given situation. In this way, even though specific images are not properly segmented or classified, it is likely that the whole set of labels is informative. Furthermore, in order to reconstruct the global situation from commonsense labels, the ideal knowledge base should exhibit two main features: *(i)* it should include a vocabulary covering a wide scope of topics, and *(ii)* it should also incorporate semantic relations between concepts.

WordNet lacks, however, semantic relations. For instance, it does not provide obvious information that a "dog" "barks". We argue that for our problem, information on contextual rather than structural relations are of greater need. ConceptNet is a semantic network designed for commonsense contextual reasoning. It is organized as a massive directed and labelled graph. It is composed by about 300,000 nodes and 1.6 million edges, corresponding to words or phrases, and relations between them, respectively. Most nodes represent common actions or chores given as phrases (e.g., "drive a car" or "buy food").

Having in mind our goal of reconstructing a global situation starting from a set of loosely related classification labels, we consider labels produced by the two stages above as a sort of *cloud of labels*. As situations change over time, the cloud of labels moves through the ConceptNet graph. Thus, the *class node* better approximating the cloud changes as well and dynamic classification can be performed by searching which class node better approximates a given cloud (i.e., labels collected within a given temporal window). The classification process is detailed below and depicted in Figure 2.

1) Each class, interesting for classification, is marked on ConceptNet as a *class node* C_i;
2) Each label coming from the previous phase is marked as a *label node* l_j. Labels that do not correspond to any node within ConceptNet are discarded;
3) For each *class node* C_i, it is computed the average shortest path pl_i to every *label node* l_j.
4) Situation classification is performed by assigning to

the time window taken into account the class C_i with the lowest pl_i.

At the end of this process, a set of images has been converted into a single ConceptNet label describing the most likely situation. This scheme, proved to be effective in dealing with *(i)* dynamism of unpredictable environments, and *(ii)* a number of noisy and misleading labels coming from automatically collected images.

III. Evaluation

In this section we describe experiments to evaluate the system we developed. We collected a small dataset using a GoPro HD camera able to collect 720p images at periodic intervals with an extremely low power consumption (around two days, with batteries fully charged, sampling every 30 seconds). We collected a dataset of 1920 images during 16 hours (not contiguous, we extracted from the whole recordings only relevant segments) of ordinary life, by sampling one image every 30 seconds. Images have been automatically timestamped and manually labelled. We took care of having at least one hundred time-contiguous images for each class we analyzed. Images of different instances of each class have been collected by three different persons. It is worth noting that being images collected from a first person perspective, overall performance do not degrade if multiple users concur in training set aquisition.

Given the dataset we collected, in each experiments we have at least one hundred images for each class against a dataset counting 1920 images overall. We considered a temporal sliding window of 300 seconds (10 images each). Neighbors search has been configured to retrieve 64 neighbors. We have chosen to retrieve 64 neighbors because the overhead introduced by searching more of them did not produce significant improvements. This evidence has been validated in [26] as well. Performance are expressed in terms of precision / recall ratio and each line averages precision and recall. Precision is defined as the fraction of retrieved instances that are relevant, while recall as the fraction of relevant instances that are retrieved. More specifically, for each image, we stored the first 16 labels associated with their confidence measure (the number of votes on Wordnet). Decreasing the minimum confidence allowed, recall increases while precision diminish. A perfect classifier is expected to achieve *precision = 1, recall = 1*. The most of the curves we plotted pass around *(precision = 0.5, recall = 0.5)*.

Although our approach is general, we focused on specific everyday life aspects that are usually relevant for scene recognition. In particular, we investigated performance in recognizing *(a)* general types of location, *(b)* domestic and *(c)* working environments, *(d)* vehicles used. To roughly quantify our contribution we compared results achieved using our approach with a modified version of [26]. Specifically, instead of segmenting images, classify each and every sub-image and identify candidate situations using ConceptNet, we searched 64 neighbors for all the images belonging to a specific temporal window and used all the retrieved labels for voting on Wordnet.

Figure 3(a) illustrates results for general scene classification. Specifically, we used the following 8 classes: *road, square, park, shop, cinema, mall, restaurant, gym*. In this experiment, the two approaches are almost comparable. The reason is that these are mainly large locations, without many obstacles. Thus the average quality of sampled images is high and details (usually emphasized through segmentation) are small compared to the scene.

Figure 3(b) shows results for domestic environments *(i.e., kitchen, living room, bathroom, bedroom, garden)*. In this experiment, our approach based on splitting the scene in a number of details and reconstructing it through commonsense knowledge outperforms the voting approach. As already discussed, the main reason behind these results is the randomness of the sampled images. If fact, while voting is effective with images with a reasonable semantic level (e.g. a good picture showing a whole living room), its performance degrades with the semantic level of considered images.

Figure 3(c), summarizes results obtained for working environments classification *(i.e., meeting room, office, corridor, leisure room)*. As for Figure 3(b), and with the same motivations, our approach outperforms voting. It is interesting to note that because of the 79 million images dataset contains less images related to these concepts, the average results of this experiment are slightly worse than for domestic environments.

Figure 3(d), finally, illustrates results for vehicle classification (i.e., bike, car, bus, train). It is important to note that we classify vehicles from a first-person perspective. The most of considered images depict what users see when they are using the vehicle. The 79 million images we use for neighbor search, instead, associate vehicles labels to images representing them from the outside. From the perspective of a user actually *watching* them instead of *using* them. Because of this reason, identifying details within images (e.g. the steering wheel, other people biking) leads to better results.

IV. Multimodal Sensor Fusion

Despite the vision sensor we presented in this paper proved to be effective for scene recognition, it is far from being perfect. Current methodologies deal with inaccurate sensors by fusing diverse sources of information. *Sensor fusion* is here used as the combining of sensory data or data derived from sensory data from disparate sources such that the resulting information is in some sense better than would be possible when these sources were used individually. The term better in this case can mean more accurate, more complete, or more dependable, or refer to the result of an emerging view.

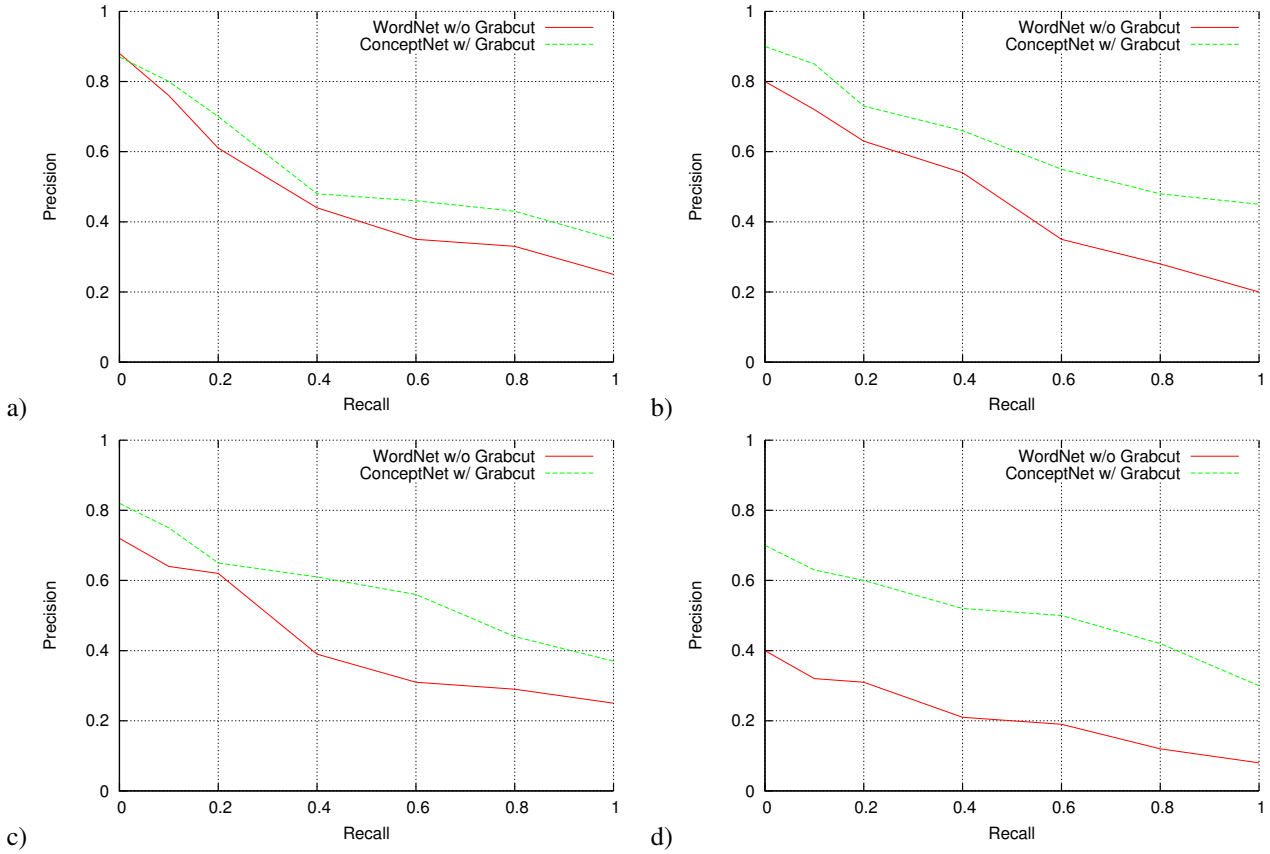

Figure 3. Precision/Recall curves for different situation recognition experiments. The system has been analyzed using a dataset comprising 1920 images collected along 16 hours (not contiguous) of ordinary sampling one image every 30 seconds. We tested ddd specific problems: *(a)* general types of location, *(b)* domestic environments, *(c)* working environments, and *(d)* vehicles used.

In this section we describe how commonsense can be used as a tool enabling sensor fusion and how it can be used to integrate the vision-based sensor presented with other sources of information. Specifically, we apply former results of our group to evaluate the benefits that can be achieved by using a vision-based location sensor in a real-world problem concerning both user activities and locations.

A. *Commonsense Multimodal Sensor Fusion*

The approach, detailed in [1], extracts well-know correlations among different facets of everyday life from a commonsense knowledge base. It can be applied to a number of cases for the sake of: *(i)* ranking classification labels produced by different classifiers on a commonsense basis (e.g., the activity classifier detects that the user is running with an high confidence and the place classifier outputs two possible labels: “park” and “swimming pool”. In this case, using commonsense, it is possible to infer that the user is more likely to be in a park that in a swimming pool); *(ii)* predicting missing labels (e.g., if a user is running but the location data is missing, “park” is a likely location).

For activities we made use of data collected from 3-axis accelerometers, sampling at 10Hz, positioned in 3 body locations (i.e., wrist, hip, ankle) and classified activities using instance-based algorithms [3]. The system has been trained to recognize 8 activities (i.e., climb, use stairs, drive, walk, read, run, use computer, stand still, drink). Classes has been described using 400 training samples each.

To classify locations we used two different sensors. The former [16], [9] samples GPS coordinates and classifies locations by querying Google Maps’ API. Specifically, this API takes as input a couple of geographic coordinates and a radius, returning a list of points of interest associated to a label coming from a predefined set. The latter, samples first-person images using a GoPro HD camera embedded into the user’s suite in front of her chest. Both of them are configured to recognize 8 locations (i.e., road, square, park, shop, cinema, mall, restaurant, gym) using a sampling period of 30s. Ground truth data have been manually annotated during 8 hours of ordinary life. The whole dataset has been split into 5 minutes long windows. The experiment consisted in automatically associate each time window to the right

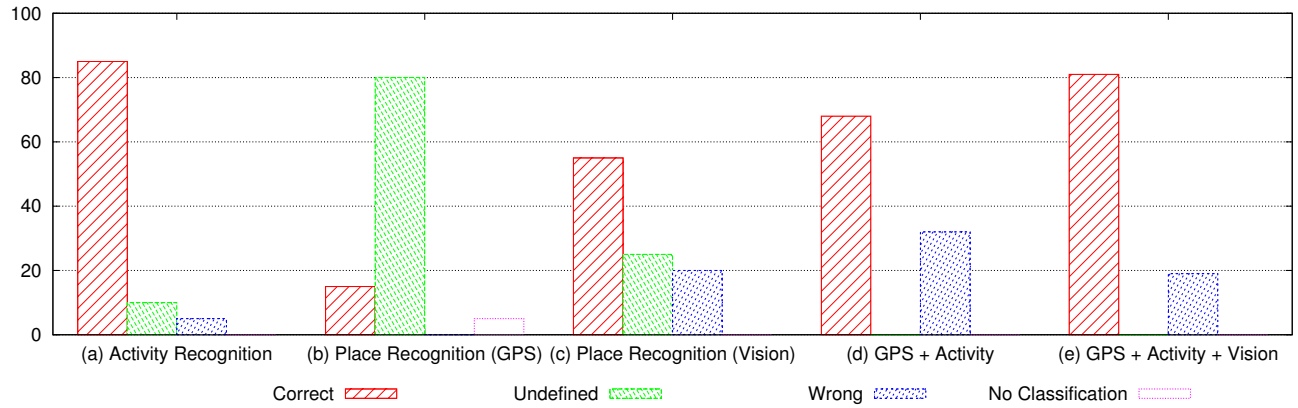

Figure 4. Experimental results. While the Activity Recognition system stand alone (a) provides reliable classifications, the Place Identification (b) can rarely provide a correct classification. Experimental results combining the Activity Recognition and Place Identification in a naive way (c) shows that there is the need for some intelligent mechanism to combine the data, indeed (d) proves that commonsense can be effectively used in this application.

(activity, location) tuple.

B. Evaluation

Experimental results have been organized according with the following four categories: *Correct Classification*: correct classification; *Undefined Classification*: more than one classification labels produced, including the correct one; *Wrong Classification*: correct classification label missing; *Missing Data*: no labels have been produced.

First, we discuss the performance of all three modules considered independently. The activity module (Figure 4a) is the most precise: around 85% of the samples are correctly classified. Location sensors provide less accurate results. However, while the vision-based sensor is able to correctly classify around 55% of the samples, the GPS-based one classifies around 80% of the samples as undefined (i.e., it returns multiple labels, including the correct one). Low accuracy of the GPS-based sensor is due to intrinsic localization errors of commercial GPS devices. To mitigate them, the system has been setup to use a search radius of 250m. Clearly, the number of reverse geo-coded locations is proportional to the search radius. The bigger the radius, the more the returned location labels. Because of this, especially in densely populated areas, the system might produce numerous false positives (i.e., undefined classifications).

Once activities and locations are transformed into structured data (i.e. classification labels), the source of information is not relevant anymore. Due to this, labels produced by all three modules can be treated in the same way and fused together on a commonsense basis using results presented in [1]. Specifically, when both location and activity labels are combined, four cases can occur: *(i)* both are available, *(ii)* only activity is available, *(iii)* only location is available, and *(iv)* no data is available. The first case allows to apply commonsense sensor fusion. In both the second and the third case, instead, commonsense can be used to identify a possible place or activity to complete the *(activity, place)* tuple.

Figures 4(d,e) show results obtained by fusing activities with locations classified by the GPS-based and vision-based sensors respectively. In both cases, the number of samples correctly classified increases while missing and undefined samples decrease to zero because of the commonsense correlations among activities and locations. As expected, vision-based location sensor produces better fused data because of its higher stand-alone accuracy.

In this experiment we demonstrated how the vision-based sensor we presented can be used in a real-world situation recognition problem and how it can be flawlessly fused with other sensors using commonsense knowledge.

V. Related Work

Many research works focus on sensor fusion at different levels, either for acquiring diverse aspects of the context or for reasoning about them to achieve a finer grained understanding or detecting inconsistencies. Early works define sensor fusion as an aggregation of logically related context data based on simple name transformation [8] either for context disambiguation [7] or complexity reduction [12]. More recent works, instead, propose to represent contextual data with a common structure in order to transparently manage diverse sources and simplify reasoning processes[5], [6]. For example, [27] proposes a model for multimodal sensors in a smart home environment that exploits the combination of multiple techniques to reduce the number of model parameters to be taken into account when a large number of sensors are used. Proact [25] combines data coming from RFIDs and an accelerometer mounted on the RFID glove in order to identify user activities. RFID tags are used to restrict the number of possible actions based on the specific

object manipulated. In [11] a system for multimodal sensor fusion specifically designed for smartphones is proposed. It uses microphone and inertial sensors data to infer user activities with light-weight bayesian learning algorithms. A similar system, also based on bayesian networks, that exploits a wider number of sensors is presented in [13]. A similar work [24] proposes healthcare applications based on bayesian networks and ontologies. An different, yet interesting, approach is presented in [2]. The core idea is to exploit trained sensors that already have a classification model in order to provide a ground truth to untrained sensors that are running learning algorithms.

On the other side, few works make use of commonsense for context recognition and classification. An interesting approach that uses commonsense to improve the activity recognition is presented in [20]. The proposed system employs models to infer user activities on the basis of the objects the user touches (as revealed by the sensing of RFID tags stick to the objects). The relations between objects and activities, is estimated using the proposed concept of Google Conditional Probability (GCP), i.e., counting the co-occurrences of the names of objects and activities across Web pages. This approach is in a similar direction of the one proposed in this paper to find relations among concept in ConceptNet . A similar approach is presented in [28], however the commonsense knowledge base has been generated by the authors, and consists of a fixed number of words for each supported action. To the best of our knowledge [16] is the only paper that tries to apply commonsense reasoning to the place identification problem. This paper uses Cyc to improve automatic place identification on the basis of the user profile, the time of the day, what happened before, etc. Both these approaches are interesting and are in the same direction of the work presented in this paper, but they limit the use of commonsense to improve a single context recognition aspect. In this paper we go further and try to exploit the commonsense to integrate different context classifiers.

To best of our knowledge there is only a work that uses commonsense to integrate different context sources. In [19] is presented a pervasive computing system to automatically discover the situation of the user by continuously overhearing his conversation. The system uses ConceptNet as the commonsense reasoning system. ConceptNet contains entities describing situations (aka. gists) and concepts related to such situations. For instance, the gist "looking for a restaurant" might have links to the "restaurant", "map", "street" and "place" concepts. The system being presented tries to spot concepts associated to gists in the users' current conversation and to classify the situation by counting the concepts. The proposed approach appears more general-purpose and can be applied to different context classifiers other than the ones proposed in the case study presented.

VI. Conclusions

Although current classifiers are still inaccurate, pervasive services often rely on multimodal classification to implement situation-recognition capabilities. In this paper we discussed the design and implementation of an innovative vision-based sensor, presented its standalone performances, and shown how it is possibly to integrate it with other well established classifiers using a commonsense knowledge base. The approach seems capable of promising classification accuracies in a number of open-ended situations.

Our future research work will include:

- Apply in-memory techniques (e.g., spectral hashing [29]) to dramatically speed up the classification process in the vision sensor, and make it usable in real-time.
- Extend out study to different kinds of sensors and to the recognition of a larger variety of situations.

Acknowledgments: work supported by the ASCENS project (EU FP7-FET, Contract No. 257414).

References

[1] N. Bicocchi, G. Castelli, M. Mamei, and F. Zambonelli. Improving situation recognition via commonsense sensor fusion. In *Proceedings of the 1st DEXA Workshop on Information Systems for Situation Awareness and Situation Management*, Toulouse (F), SEP 2011. IEEE CS Press.

[2] N. Bicocchi, M. Mamei, A. Prati, R. Cucchiara, and F. Zambonelli. Pervasive self-learning with multi-modal distributed sensors. In *Proceedings of the 2008 Second IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pages 61–66, Washington, DC, USA, 2008. IEEE CS Press.

[3] N. Bicocchi, M. Mamei, and F. Zambonelli. Detecting activities from body-worn accelerometers via instance-based algorithms. *Pervasive and Mobile Computing*, 6(4):482–495, 2010.

[4] C. A. Boano, M. Lasagni, K. Römer, and T. Lange. Accurate Temperature Measurements for Medical Research using Body Sensor Networks. In *Proceedings of the 2nd International Workshop on Self-Organizing Real-Time Systems (SORT)*, pages 189–198, Newport Beach, CA, USA, Mar. 2011. IEEE CS Press.

[5] J. Bravo, R. Hervás, G. Chavira, and S. W. Nava. Modeling contexts by rfid-sensor fusion. In *IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 30–34, Los Alamitos, CA, USA, 2006. IEEE CS Press.

[6] G. Castelli, A. Rosi, M. Mamei, and F. Zambonelli. A simple model and infrastructure for context-aware browsing of the world. In *IEEE International Conference on Pervasive Computing and Communications*,

pages 229–238, Los Alamitos, CA, USA, 2007. IEEE CS Press.

[7] A. Dey, J. Mankoff, G. Abowd, and S. Carter. Distributed mediation of ambiguous context in aware environments. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*, UIST '02, pages 121–130, New York, NY, USA, 2002. ACM.

[8] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction*, 16:97–166, December 2001.

[9] L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *Proceedings of 11th IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 397–402. IEEE CS Press, 2011.

[10] C. Galleguillos and S. J. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.

[11] R. K. Ganti, S. Srinivasan, and A. Gacic. Multisensor fusion in smartphones for lifestyle monitoring. In *Proceedings of the International Conference on Body Sensor Networks*, pages 36–43. IEEE CS Press, 2010.

[12] J. I. Hong and J. A. Landay. An infrastructure approach to context-aware computing. *Human-Computer Interaction*, 16:287–303, December 2001.

[13] K. S. Hwang and S. B. Cho. Life log management based on machine learning technique. In *Proocedings of International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 691–696. IEEE CS Press, 2008.

[14] E. Kim, S. Helal, and D. Cook. Human activity recognition and pattern discovery. *Pervasive Computing, IEEE*, 9(1):48 –53, January 2010.

[15] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

[16] M. Mamei. Applying commonsense reasoning to place identification. *IJHCR*, 1(2):36–53, 2010.

[17] A. Oliva. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41:176–210, 2000.

[18] J. Park and H. Y. Song. Multilevel localization for mobile sensor network platforms. In *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, pages 711–718, October 2008.

[19] A. Pentland, T. Choudhury, N. Eagle, and P. Singh. Human dynamics: computation for organizations. *Pattern Recognition Letters*, 26:503–511, 2005.

[20] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, 2004.

[21] M. Pijl, S. van de Par, and C. Shan. An event-based approach to multi-modal activity modeling and recognition. In *Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on*, pages 98 –106, April 2010.

[22] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 413–420, 2009.

[23] C. Rother, V. Kolmogorov, and A. Blake. "proceedings": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23:309–314, August 2004.

[24] N. Roy, T. Gu, and S. K. Das. Supporting pervasive computing applications with active context fusion and semantic context delivery. *Pervasive and Mobile Computing*, 6(1):21 – 42, 2010.

[25] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele. Adl recognition based on the combination of rfid and accelerometer sensing. In *2nd International Conference on Pervasive Computing Technologies for Healthcare*, pages 258–263, 2008.

[26] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 30:1958–1970, November 2008.

[27] D. T. Tran and D. Q. Phung. A probabilistic model with parsinomious representation for sensor fusion in recognizing activity in pervasive environment. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, ICPR '06, pages 168–172, Washington, DC, USA, 2006. IEEE CS Press.

[28] S. Wang, W. Pentney, A.-M. Popescu, T. Choudhury, and M. Philipose. Common sense based joint training of human activity recognizers. In *Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2237–2242, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[29] Y. Weiss, A. B. Torralba, and R. Fergus. Spectral hashing. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, pages 1753–1760. MIT Press, 2008.

[30] Widyawan, G. Pirkl, D. Munaretto, C. Fischer, C. An, P. Lukowicz, M. Klepal, A. Timm-Giel, J. Widmer, D. Pesch, and H. Gellersen. Virtual lifeline: Multimodal sensor data fusion for robust navigation in unknown environments. *Pervasive and Mobile Computing*, 2011.

[31] C. Zhu and W. Sheng. Recognizing human daily activity using a single inertial sensor. In *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, pages 282 –287, July 2010.