Improving Situation Recognition via Commonsense Sensor Fusion

Nicola Bicocchi
Dipartimento di Ingegneria dell’Informazione
Università di Modena e Reggio Emilia
Modena, Italia
name.surname@unimore.it

Gabriella Castelli, Marco Mamei, Franco Zambonelli
Dipartimento di Scienze e Metodi dell’Ingegneria
Università di Modena e Reggio Emilia
Reggio Emilia, Italia
name.surname@unimore.it

Abstract—Pervasive services often rely on multi-modal classification to implement situation-recognition capabilities. However, current classifiers are still inaccurate and unreliable. In this paper we present preliminary results obtained with a novel approach that combines well established classifiers using a commonsense knowledge base. The approach maps classification labels produced by independent classifiers to concepts organized within the ConceptNet network. Then it verifies their semantic proximity by implementing a greedy approximate sub-graph search algorithm. Specifically, different classifiers are fused together on a commonsense basis for both: (i) improve classification accuracy and (ii) deal with missing labels. Experimental results are discussed through a real-world case study in which two classifiers are fused to recognize both user’s activities and visited locations.

Keywords-Pervasive Computing; Activity Recognition; Mobility; Commonsense Knowledge

I. INTRODUCTION

Pervasive services need to perceive and understand their operating environment. They have been conceived to overcome several limitations of previous human-machine interaction mechanisms. The crisp and rigid behavior of traditional computers, driven by explicit interactions such as pressing keyboard’s keys, is leaving room to adaptable behaviors driven by machines’ perception of their operating environment and users’ needs. Many services, analyzing different facets of our world, frequently cooperate to provide services with a coherent representation of the environment. However, despite the many facets of our life are strictly tied from the practical viewpoint (E.g., if a user is running he is likely to be in suitable location such as a park or a gym), it is difficult to exploit their correlation using traditional learning techniques (E.g., bagging, boosting) [7], [2]. On the other side, treating each facet as an independent variable might lead to unrealistic results. For instance, locations and activities are strictly correlated. Thus, relying on two separate classifiers might be undesirable.

In this paper we tackle the problem of enabling situation-recognition capabilities by fusing different sensor contributions. Specifically, we propose to extract well-know correlations among different facets of everyday life from a commonsense knowledge base. The approach is general and can be applied to a number of cases involving commonsense for the sake of: (i) ranking classification labels produced by different classifiers on a commonsense basis (e.g., the action 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, it is possible to propose “park” as a likely location).

More in details, the paper contains the following contributions and insights: (i) it describes a greedy search algorithm to measure the semantic proximity of two concepts within the ConceptNet [8] network; and (ii) it applies the proposed algorithm to a specific sensor fusion problem involving user’s locations and activities.

Accordingly, the rest of the paper is organized as follows: Section II formally defines the problem of commonsense sensor fusion and describes the proposed algorithm. Section III describes the experimental testbed we implemented to validate our proposal. Section IV details experimental results under different configurations. Section V discusses related work. Finally, Section VI concludes the paper.

II. COMMONSENSE SENSOR FUSION

The proposed approach is based on the assumption that commonsense knowledge can be used to measure the semantic proximity among concepts. The more two concepts are semantically proximate, the more it is likely they have been recognized within the same situation [9]. In this section we formally introduce the approach.

A. Problem Definition

Let us consider a set of \( n \) classifiers \( C_1, \ldots, C_n \), each one delegated to recognize a specific facet of the environment. Each classifier is able to deal with uncertainties by producing \( m \) labels \( l_1(C_x, t), \ldots, l_m(C_x, t) \) for each data sample. Given that, the overall perception of the environment can be represented as a tuple \((l_1(C_1, t), \ldots, l_m(C_1, t)), \ldots, (l_1(C_n, t), \ldots, l_m(C_n, t))\).
A preliminary round of experiments with ConceptNet led us to identify the following principles:

1) Proximity increases with the number of unique paths. However, this is not a reliable indicator given that even completely unrelated concepts might be connected through long paths or highly connected nodes;
2) Proximity decreases with the length of the shortest path; nodes connected directly or through some niche edges are in a short distance, hence they are proximate;
3) Connections going through highly connected nodes increase ambiguity, therefore proximity should be inversely proportional to the degrees of visited nodes;
4) ConceptNet has been created from natural-language assertions. Thus, errors are frequent and algorithms have to be noise-tolerant;
5) Computational complexity should be low.

According to the above considerations, given a knowledge base represented as an oriented graph \( G = (V, E) \), two nodes \( u \) and \( v \) connected both by an edge \( e_{u,v} \), and a multi-hop path \( p_{u,v} \), we define:

(i) the weight \( w(u) \) of node \( u \) as its in-degree;
(ii) the weight \( w(l_{u,v}) \) of an edge connecting \( u \) and \( v \) as \( \max(w(u), w(v)) \); and
(iii) the weight \( w(p_{u,v}) \) of a path from \( u \) to \( v \) as the sum of the weight of all the edges composing \( p_{u,v} \).

To establish which tuple contains the most proximate concepts on a commonsense basis?

### B. Semantic Proximity over ConceptNet

First, we define the most relevant requirements for our commonsense knowledge base. Since we are analyzing context recognition problems, 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 tricky relations clear to humans but hard to discover in an automatic way.

ConceptNet best suits these requirements. It is a semantic network designed for commonsense contextual reasoning. It provides commonsense contextual associations not offered by any other knowledge base. ConceptNet is organized as a massive directed and labelled graph. It is made of 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., “ride a bike” or “play soccer”). It contains over 86,000 leaf nodes and approximately 25,000 root nodes. The average degree of the network is approximately 4.7.

In this paper, we tackle the problem of ranking all the possible tuples provided by \( n \) classifiers on a commonsense basis.

The general problem of commonsense tuple ranking can be expressed, without loss of generality, in this way: given 2 tuples both composed by commonsense concepts, \((l_1(C_1, t), l_1(C_2, t))\) and \((l_2(C_1, t), l_2(C_2, t))\), is it possible to establish which tuple contains the most proximate concepts on a commonsense basis?

### III. A Case Study

To assess the relevance of our ideas, we used a specific instance of the general problem. We prototyped an automatic diary able to recognize both activities performed by the user and locations she visited. Activities have been recognized from accelerometer data while locations from GPS traces. Both modules have been configured to produce multiple labels to deal with uncertainties. In these cases, commonsense reasoning is applied.

#### A. Activity Recognition

To classify user’s activities we made use of the system detailed in [4]. It collects data from 3-axis accelerometers, sampling at 10Hz, positioned in 3 body locations (i.e., wrist,
Main() :
  \(i \leftarrow \text{findPathWeights}(u, v)\)
  \(j \leftarrow \text{findPathWeights}(p, q)\)
  \(W \leftarrow \min(\min(i), \min(j))\)
  \(S_i \leftarrow \text{normalize}(i, W)\)
  \(S_j \leftarrow \text{normalize}(j, W)\)
  \[\text{if } S_i \geq S_j \text{ then} \]
  \[\text{return } (u, v)\]
  \[\text{end if}\]
  \[\text{return } (p, q)\]

\(\text{Normalize(weights, minWeight)} :\)
\(S \leftarrow 0\)
\[\text{for each } w \text{ in weights do}\]
\(w \leftarrow 1 - \left(\frac{w}{\text{minWeight}} - 1\right)\)
\[\text{if } w \geq 0 \text{ then}\]
\(S \leftarrow S + w\)
\[\text{end if}\]
\[\text{end for}\]

Figure 2. Semantic proximity algorithm. Each distinct path connecting \(u, v\) and \(p, q\) have to be found. Then each path weight is normalized and long paths are discarded. Finally, the weights of all the remaining paths are aggregated, and the two nodes associated with the highest value are returned.

...and stores the result. This way, it is possible to discard raw samplings and store only high-level activity labels, allowing the execution of 4+ hours experiments without using heavy and obtrusive equipment.

2) Second, we modified it to deal with uncertainties. Instead of producing a single label for each sensor sampling, we implemented a mechanism to produce multiple labels associated with a degree of confidence. Specifically, for each sample to be classified, \(k\) nearest neighbors (associated to \(q\) classes, \(k = 64, q \leq k\)) are identified. The sample is then associated to all the classes (at most 3) associated to all the training samples.

**B. Location Recognition**

To classify user’s location we implemented a tool [10], [5] for Symbian OS. It samples GPS coordinates and classifies user’s location by querying the reverse geocoding Google Maps’ API. Specifically, this API takes in 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 (i.e., road, square, park, shop, cinema, mall, restaurant, gym). Unfortunately several practical drawbacks affect this process:

1) Smart phones are not equipped with high-precision GPS receivers. Under normal operating conditions this error is smaller than 100m [10]. However, whenever the GPS signal is not received perfectly (e.g., users moving back and forth around large buildings, foggy weather, etc.), the error can reach 250m.

2) Google Maps database is not perfect. Although we do not have accurate statistics, we noticed that a portion of locations is still missing. Furthermore, locations’ coordinates are not always precise. Finally, Google Maps does not provide information about locations’ geometry. Due to this, especially with large-sized instances (e.g., parks, squares) locations can be misclassified. For example, a user running close to the border of a park is likely to be associated to the shops she is facing instead of to the park itself.

To mitigate these problems and avoid false negatives, 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. To reduce them, while keeping an acceptable level of false negatives, we implemented 3 filters acting on different dimensions of the GPS signal. Specifically:

- **Date/Time.** It acts on the assumption that each class of locations (i.e., a label) is more likely to be visited during defined portions of the week. Each label associated with a probability lower than a certain threshold is filtered out.

- **Signal Interruption.** It acts on the assumption that each class of locations is fairly characterized by the duration of the visit. This duration is usually related with a GPS signal interruption. Each label associated with a probability lower than a certain threshold is filtered out.

- **Speed.** A common misclassification happens when a user moving on a street and is associated to all the locations she goes by. It filters out each label not compatible with measured speed.

**C. ConceptNet**

Both modules described in this section periodically produce a tuple of classification labels. A third module, imple-
menting the algorithm described in Section II, analyzes their outputs attempting to identify the most probable \((activity, location)\) tuple. From another perspective, and considering the higher accuracy of the activity module, commonsense can be seen as an additional multi-modal filtering mechanism integrated within the localization module. Experimental results are explained and discussed in the next section.

IV. EXPERIMENTAL RESULTS

A volunteer, equipped with the system described in Section III, collected data while going about his normal life and manually annotating ground truth data.

The activity recognition module, has been trained to recognize 8 activities (i.e., climb, use stairs, drive, walk, read, run, use computer, stand still, drink). For each class, 300 training samples have been selected. The location identification module, instead, has been run on a Nokia N95 smartphone using a sampling period of 30 seconds. GPS coordinates has been associated with a label coming a predefined set (i.e., street, university, bar, park and library).

Experimental results have been organized according with the following categories:

- **Correct Classification.** Both the activity recognition and place identification modules provide a single and correct result that can be easily put together;
- **Undefined Classification.** One or both the classifiers provide more that a results, in this case there is the need for an additional mechanism to choose the most proper activity and/or location;
- **Wrong Classification.** A classifier or both of them provided a wrong classification;
- **Missing Data.** A classifier or both of them didn’t provide data to be classified.

We first discuss the performance of both modules, considered independently. The activity module produces reliable classifications (see Figure 3(a)). Around 90% of the samples are correctly classified, while 9% are undefined and only the remaining 1% are wrong. The location module, instead, is less accurate. Around 80% of the samples are classified as undefined because multiple labels are produced (see Figure 3(b)). Indeed the place identification module can filter out some labels but still needs additional information to improve accuracy.

When both location and activity labels are combined using ConceptNet, 4 cases can occur: (i) both are available, (ii) only activity is available, (iii) only location is available, (iv) no data 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. Finally, the fourth case does not allow to use commonsense. However, different processing based on predictions can be performed.

Figures 3(c,d) show the results of the combined classifier and compares them with a basic system without commonsense integration. This system (see Figure 3(c)) classifies the majority of samples as Unclassified, mainly due to the place identification module’s output.

The combined system (see Figure 3(d)), instead, shows a significant improvement with 75% of data correctly classified and 25% of wrong classifications. It is worth noticing that the Undefined Classification category is lowered to zero meaning that ConceptNet is always capable of providing a ranking of action-place couples. Furthermore, the Missing data category is lowered to zero, in fact one of the advantage of the use of ConceptNet is to provide missing data. Please note that in our experiment we never experienced the concurrent lack of both sensorial data, that should have called for different strategies similar to activity and location prediction, such as bayesian networks [3].

V. RELATED WORK

Many works focus on data fusion at different levels, either for acquiring and make accessible diverse contextual aspects or for reasoning about them. The traditional approach make use of probabilistic models. Proact [13] combines data coming from RFID’s and an accelerometer mounted on the RFID glove in order to identify activities, the RFID tags objects and are used to restrict the number of possible actions based on the specific object manipulated. In [6] a system for multi-modal sensor fusion specifically designed for smartphone is proposed. The system exploits data coming from the microphone and inertial sensors on the mobile for inferring high level activities with light-weight bayesian learning algorithms.

Few works make use of commonsense for situation recognition. An interesting approach is presented in [12]. It uses RFID to trace a set of everyday objects and infers user activities by making use of Google searches. Alternatively, [10] applies commonsense to localization. It uses Cyc to improve automatic place identification on the basis of user historical data. However, both these approaches limit the use of commonsense to improve a single contextual aspect. Alternatively, in this paper we use commonsense to integrate multiple aspects.

To best of our knowledge there is only a work that uses commonsense to integrate different context sources. Pentland et. al. [11] presented a user-centric situation recognition system able to continuously overhearing users’ conversations and differerusing ConceptNet as reasoning system. The approach we propose in this paper appears more general-purpose and can be applied to different context classifiers other than the ones proposed in the evaluated case study.

VI. CONCLUSIONS

Although pervasive services often rely on situation-recognition capabilities, current classifiers are still inaccurate
and unreliable. In this paper we presented preliminary results we obtained with a novel approach that combines well established classifiers using the ConceptNet knowledge base. Different classifiers are fused together on a commonsense basis for both: (i) improve classification accuracy and (ii) deal with missing labels. The approach has been discussed through a realistic case study focused on the recognition of both locations visited and activities performed by a user. Results have been encouraging and apparently indicates that our approach can be applied to different scenarios.

ACKNOWLEDGMENT

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

REFERENCES