Experiences on Sensor Fusion with Commonsense Reasoning

Nicola Bicocchi[†], Gabriella Castelli[‡], Matteo Lasagni[*‡], Marco Mamei[‡], Franco Zambonelli[‡]

[†]*Dip. di Ingegneria dell'Informazione*
Universitá di Modena e Reggio Emilia
Modena, Italia
name.surname@unimore.it

[*]*Institute of Computer Engineering*
University of Lübeck
Lübeck, Germany
lasagni@iti.uni-luebeck.de

[‡]*Dip. di Scienze e Metodi dell'Ingegneria*
Universitá di Modena e Reggio Emilia
Reggio Emilia, Italia
name.surname@unimore.it

***Abstract*—Multi-modal sensor fusion recently became a widespread technique to provide pervasive services with context-recognition capabilities. However, classifiers commonly used to implement this technique are still far from being perfect. Thus, fusion algorithms able to deal with significant inaccuracies are required. In this paper we present preliminary results obtained with a novel approach that combines diverse classifiers through commonsense reasoning. The approach maps classification labels produced by classifiers to concepts organized within the ConceptNet network. Then it verifies their semantic proximity by implementing a greedy sub-graph search algorithm. Specifically, different classifiers are fused together on a commonsense basis for both: *(i)* improving classification accuracy and *(ii)* dealing with missing labels. Experimental results are discussed through a real-world case study in which three classifiers are fused to recognize both user activities and locations.**

***Keywords*-Pervasive Computing; Localization; Mobility; Commonsense Knowledge; Image recognition; Activity Recognition;**

I. Introduction

The automatic and unobtrusive identification of user's activities and situation from sensor data is one of the key goals of context-aware computing. Several opportunities can arise from such a situation identification in that pervasive computing services could become capable of dynamically adapting their behavior to the current situation.

While advances in hardware technologies are making feasible to collect a information about several facet of the user situation, it is still difficult to organize and aggregate all the collected information in a coherent, expressive and semantically-rich representation. From a more technical viewpoint, 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 [6], [1]. 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 context-recognition capabilities by fusing different sensor contributions. Specifically, we propose to extract well-known 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 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 rather than 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 detail, the paper contains the following contributions and insights:

1) it describes a greedy search algorithm to measure the semantic proximity of two concepts within the ConceptNet [7] network;
2) it applies the proposed algorithm to a specific sensor fusion problem involving both user location and activity.

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 highlights experimental results obtained through a realistic case study. Finally, Section V 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 proximate, the more it is likely they have been recognized within the same context [8]. In this section we formally introduce the approach.

A. Problem Definition

Let us consider a set of n classifiers $C_1..C_n$, each one delegated to recognize a specific facet of the environment. Each classifier C_x is able to deal with uncertainties by producing (at every time step t) m labels $l_1(C_x,t),...,l_m(C_x,t)$ for each data sample. Given that, the overall perception of the environment can be represented as a tuple $((l_1(C_1,t),...,l_m(C_1,t)),...,(l_1(C_n,t),...,l_m(C_n,t)))$.

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?

Measuring commonsense proximity requires two key conditions to be met. In particular: *(i)* a knowledge base containing both a vocabulary covering a wide scope of topics and semantic relations hard to be discovered in an automatic way; and *(ii)* an algorithm for computing semantic proximity.

B. ConceptNet

The first condition is best addressed by ConceptNet. It is a semantic network designed for commonsense contextual reasoning. It was automatically built from a collection of 700,000 sentences, a corpus being a result of collaboration of some 14,000 people. 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., "drive a car" or "buy food"). Its structure is uneven, with a group of highly connected nodes, and "person" being the most connected, having in- degree of about 30,000 and out- degree of over 50,000. There are over 86,000 leaf nodes and approximately 25,000 root nodes. The average degree of the network is approximately 4.7.

C. Semantic Proximity

To meet the second requirement, we started from a preliminary round of experiments with ConceptNet that led us to 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;

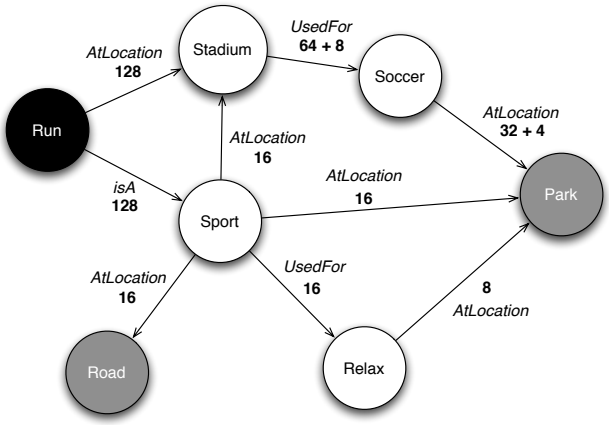

Figure 1. Concept proximity algorithm in action. 256 units of substance s are injected into node Run. Then, the substance diffuses over the graph and halves by evaporation ($\alpha = 0.5$) at each node it visits. The amounts of s that reach nodes Park and Road are 60 and 16 respectively. Park is considered more proximate than Road to Run.

Majewski et al. recently proposed an interesting algorithm for commonsense text categorization inspired by similar observations [8]. Despite having been conceived for a different problem, it can be applied to localization as well. The algorithm is based on the assumption that proximity among concepts is proportional to the amount of some substance s that reaches the destination node v as a result of injection to node u. The procedure has been built around two key biological paradigms such as *diffusion* and *evaporation* and works as follow:

1) a given amount of substance s is injected to a node u;
2) at every node, a fraction α of the substance evaporates and leaves the node;
3) at every node, the substance diffuses into smaller flows proportional to the out degree of the node;
4) nodes never overflow. If multiple paths visit the same node, the previous amount of substance s can be incremented;
5) target nodes are ranked according to the amount of substance s received.

Figure 1 exemplifyies the algorithm in action. A certain amount (i.e., 256 units) of substance s is injected into a node (i.e., *Run*). Then, the substance diffuses over the graph and halves by evaporation at each node it visits. The amounts of s that reach nodes *Park* and *Road* are 60 and 16 respectively. *Park* is considered more proximate than *Road* to *Run*.

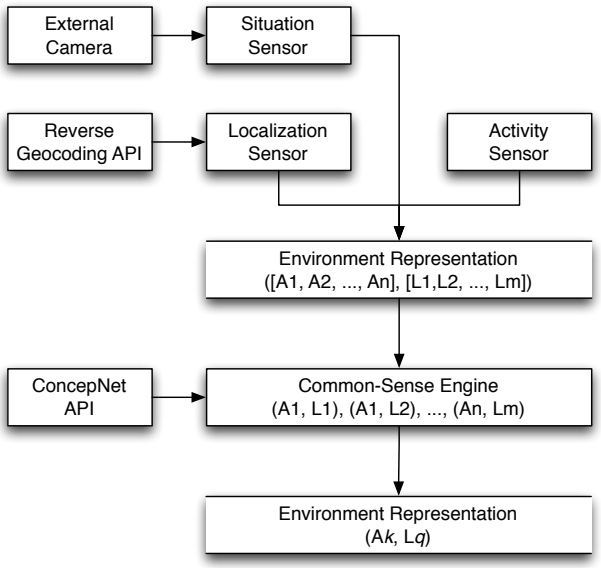

Figure 2. Tuples containing labels coming from three different sensors are ranked by making use of the ConceptNet knowledge base.

It is worth noticing that this approach can easily handle the fact that different classifiers might produce the same set of labels (i.e., classifiers observing the same facets of reality). In fact, if a label compares multiple times it is sufficient to multiply the amount of substance injected into the corresponding nodes. Furthermore, this approach permits to assign different weights to different classifiers is a straightforward way.

Finally, it is interesting to note how this algorithm matches with the principles we deduced from our preliminary studies on ConceptNet. In fact: *(i)* the evaporation process assures that short paths imply high proximity; while *(ii)* the diffusion process takes into account the total amount of connections among two concepts while diminishing the relevance of highly-connected paths.

III. Sensors setup and data acquisition

To assess our ideas we tackled the problem of automatically inferring both locations visited and activities performed by a mobile user. We made use of three modules: *(i)* a GPS-based localization sensor, *(ii)* a vision-based ambient recognition sensor, and *(iii)* an accelerometer-based activity sensor. As shown in Figure 2, their classification labels are fed to our commonsense engine for both improving classification accuracy and dealing with missing labels. Nevertheless, our results abstract from specific implementation details and can be reproduced with alternative sensors with comparable precision and recall. It is worth noticing that experiences described in this paper could be significant because the sensors used are likely to be widespread in the next few years.

A. Localization Sensor

This sensor [4] samples GPS coordinates and performs localization by querying Google Maps. Specifically, Google Maps takes as input a couple of geographic coordinates and a radius, returning a list of locations of interest associated with a label coming from a predefined set. Unfortunately, two principal drawbacks affect this process.

First, smart phones are not equipped with high-precision GPS receivers. Under normal operating conditions this error is smaller than 100m [9]. However, whenever the GPS signal is not received perfectly, the error can reach 200m.

Second, Google Maps database is not perfect. Although we do not have accurate statistics, we noticed that a portion of locations is still missing and coordinates might be unprecise. Furthermore, 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 issues while avoiding false negatives, a search radius of 250m has been used. Clearly, the number of reverse geo-coded locations is proportional to the search radius. Because of this, especially in urban areas, the system might produce numerous false positives. To reduce them, we implemented three filters acting on different features of the GPS signal.

DateTime filter acts on the assumption that each label is more likely to be visited during defined portions of the week. Thus, we associated to each label a probability distribution (i.e., 24-7) describing how likely that category of places is going to be visited. Every label associated with a probability lower than a certain threshold is filtered out.

Speed filter works on continuous GPS signals (suggesting an outdoor location). A common misclassification happens when a user moving on a street an is associated to locations she goes by. To avoid this, the filter analyzes user's speed. If the user is moving, only *road*, *park* and *square* are allowed.

Finally, Interruption filter works on discontinuous GPS signals indicating, with high probability, an indoor location. It works on the assumption that each category of places is fairly characterized by the duration of the visit. Thus, we defined for each category a probability distribution of durations. Each label associated with a probability lower than a certain threshold is filtered out.

B. Vision-based Ambient Recognition Sensor

Low-cost 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 context recognition.

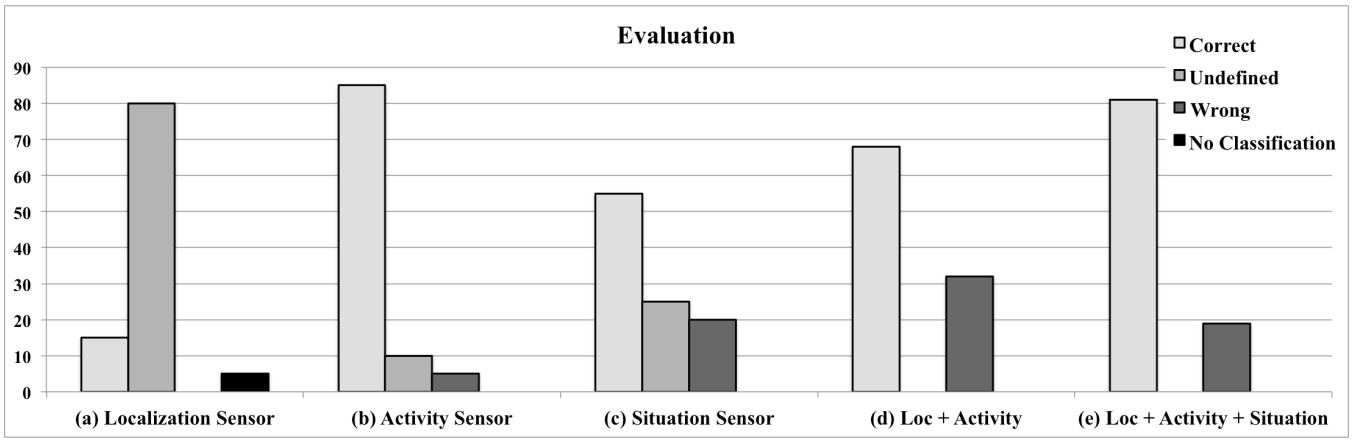

Figure 3. $Correct$ refers to correctly provided classification labels. $Undefined$ accounts cases in which classifiers provide more than one result. $Wrong$ refers to incorrectly provided classification labels. $MissingData$ accounts cases in which classifiers did not provide any result. Localization Sensor(a) rarely provides correct classifications, while the activity(b) and ambient(c) sensors correctly classify around 80% and 50% of the samples respectively. Figures (d) (e) show the benefits achievable by making use of commonsense to fuse GPS-based localization with user activities and ambients provided by other sensors.

Results presented in [12], showed that it is possible to recognize objects, persons and situations using large datasets combined with non-parametric algorithms. The sensor we implemented [2] classifies a set of images (i.e., the temporal window in which images have been captured) through the following workflow:

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 by searching its nearest neighbors within a loosely labelled dataset of $7.9x10^7$ images. 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.
3) *Situation recognition*: labels assigned to images (and segmented sub-images) taken within a temporal window are positioned within the ConceptNet knowledge base as *label nodes*. Classification is performed by searching *situation nodes* with the shortest average path to the individual *label nodes* so as to eventually identify a label that re-construct a global comprehensive situation.

As a result, a simple camera can become an effective ambient sensor, based on which users can compactly classify situations around them. This approach, relying on massive dataset, allows to recognize a wide spectrum of classes. However, for the experiments described here, we configured it to recognize the same set of ambients returned by the localization sensor (i.e., road, square, park, shop, cinema, mall, restaurant, gym).

C. Activity Recognition Sensor

To classify user's activities we made use of the system detailed in [3]. It collects data from 3-axis accelerometers, sampling at 10Hz, positioned in 3 body locations (i.e., wrist, hip, ankle) and classifies activities using instance-based algorithms. Furthermore, considering that human activities have a minimum duration, it aggregates classification results over a sliding window and performs majority voting on that window. Each window is associated with the most frequent label. For the sake of the experimentation, we modified this module in two ways:

First, we implemented both training and classification modules on Sun Spot nodes. Instance-based algorithm perfectly suit this need in that they support on-line classification and training and can be implemented on resource-constrained devices. Client nodes send their samplings to a master node which classifies them 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.

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, $k = 64$, $q <= k$) are identified. The sample is then associated to all the classes (at most 3) associated to al least $k/2q$ training samples.

IV. Experimental Results

In this section we describe experiments and discuss results. The activity sensor, sampling at 10Hz, has been trained to recognize 8 classes (i.e., climb, use stairs, drive, walk, read, run, use computer, stand still, drink) described by 400 training samples each.

Absolute localization is provided by the GPS signal, while both localization and ambient recognition sensors deal with situation recognition. Both of them have been configured to recognize 8 location classes (i.e., road, square, park, shop, cinema, mall, restaurant, gym). A volunteer equipped with the sensing system collected data while going about his ordinary life and manually annotated the ground truth, the experiment duration has been 8 hours and the sampling period 30s. The whole dataset has been split into 5 minutes long windows, to each of them a *(activity, location)* tuple has been associated.

First, we discuss the performance of all three modules considered independently. Localization sensor exhibits the worst performance: around 80% of the samples are classified as undefined (i.e., multiple labels, including the correct one, are returned). The activity sensor (Figure 3a), instead, is the most precise: around 85% of the samples are correctly classified. Finally, the vision-based sensor is able to correctly classify around 50% of the samples, equally dividing the rest between wrong and undefined classifications.

Once sensed data are classified, the system deals with labels that are fused together on a commonsense basis regardless of the data sources. In particular, when combining location and activity labels, 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 to improve the sensor fusion. In both the second and the third case, instead, commonsense can be used to identify a candidate place or activity to complete the *(activity, place)* tuple.

Figures 3(d,e) show results obtained by fusing contributions coming from multiple sensors. In both cases, the number of samples correctly classified increases and missing and undefined samples are lowered to zero because of commonsense relations. As expected, introducing vision-based sensor 3(e) further increases accuracy.

The experimental results proves that commonsense knowledge can be fruitfully used in real-world sensor fusion problems to seamlessly integrate diverse sources of information enabling multi-modal recognition. Furthermore, it is worth noticing that, due to the wide spectrum of concepts organized within ConceptNet, the same approach is suitable to be applied in a number of different scenarios.

V. Conclusions

Although pervasive services require to perceive (i.e., classify) their operating environment to generate their context-models, classifiers commonly employed to carry out this task are still inaccurate and unreliable. In this paper we presented a novel approach that combines diverse pervasive sensors using the ConceptNet knowledge base. User activities and their commonsense relations with location classes have been used to improve classification accuracy. Experiments have been discussed through a realistic case study involving a GPS-based location sensor, a vision-based situation recognition sensor and an accelerometer-based activity sensor.

VI. Related Work

Several works are related to this paper, in this section after referring some key papers for sensor fusion, we mainly focus in particular on sensor fusion and the use common sense for sensorial data merging.

The traditional approach make use of probabilistic models. Proact [11] combines data coming from RFIDs 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 [5] 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.

To best of our knowledge there are only few approaches that make use of commonsense for context recognition and classification. In [10] the sensing of RFID tags sticked to everyday objects is exploited to infer user activities by making use of a probabilistic algorithm that is based on Google searches. This approach is in a similar direction of the one proposed in this paper to find relations among concept in ConceptNet. In [13] a similar approach is introduced. However, the commonsense knowledge base has been generated by the authors thus simplifying a number of related issues. We found that [9] is the only paper that tries to apply common sense 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 common sense to improve a single context recognition aspect. In this paper we go further and try to exploit the common sense to integrate different context classifiers.

To best of our knowledge there is only a work that uses commonsense to integrate different context sources. Pentland in [11] presented a user-centric situation recognition system able to continuously overhearing users conversations exploiting 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.

ACKNOWLEDGMENT

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

REFERENCES

[1] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni. A survey of context modelling and reasoning techniques. *Pervasive Mobile Computing*, 6:161–180, April 2010.

[2] N. Bicocchi, M. Lasagni, and F. Zambonelli. Bridging vision and commonsense for multimodal situation recognition in pervasive systems. In *International Conference on Pervasive Computing and Communications*, Lugano, Switzerland, 2012.

[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] L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *IEEE International Conference on Pervasive Computing and Communications, Workshop on Context Modeling and Reasoning*, Seattle, WA, USA, 2011.

[5] R. K. Ganti, S. Srinivasan, and A. Gacic. Multisensor fusion in smartphones for lifestyle monitoring. In *Proceedings of the 2010 International Conference on Body Sensor Networks*, BSN '10, pages 36–43, Washington, DC, USA, 2010. IEEE Computer Society.

[6] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Comparing boosting and bagging techniques with noisy and imbalanced data. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, PP(99):1 –17, 2010.

[7] H. Liu and P. Singh. Conceptnet, a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226, October 2004.

[8] P. Majewski and J. Szymański. Neural information processing. chapter Text Categorization with Semantic Commonsense Knowledge: First Results, pages 769–778. Springer-Verlag, Berlin, Heidelberg, 2008.

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

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

[11] M. Stikic, T. Huynh, K. Van Laerhoven, and B. Schiele. Adl recognition based on the combination of rfid and accelerometer sensing. In *Second International Conference on Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008.*, 30 2008.

[12] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1958–1970, November 2008.

[13] 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.