Learning Situation Models for Providing Context-Aware Services

Oliver Brdiczka, James L. Crowley, Patrick Reignier

Learning Situation Models for Providing Context-Aware Services

O. Brdiczka, J. L. Crowley, P. Reignier

Project PRIMA
INRIA Rhône-Alpes
Montbonnot, France
{brdiczka, crowley, reignier}@inrialpes.fr

In order to provide information and communication services without disrupting human activity, information services must implicitly conform to the current context of human activity. However, the variability of human environments and human preferences make it impossible to preprogram the appropriate behaviors for a context aware service. One approach to overcoming this obstacle is to have services adapt behavior to individual preferences though feedback from users. This article describes a method for learning situation models to drive context-aware services. With this approach an initial simplified situation model is adapted to accommodate user preferences by a supervised learning algorithm using feedback from users. To bootstrap this process, the initial situation model is acquired by applying an automatic segmentation process to sample observation of human activities. This model is subsequently adapted to different operating environments and human preferences through interaction with users, using a supervised learning algorithm.

Introduction

Research in intelligent environments seeks to provide an enabling technology for a variety of new services. The key to acceptability for such services is unobtrusive behavior. Services that create distractions and unwanted interruptions are unlikely to find acceptance. Services must conform to human social conventions, and comply with rules for polite interaction. Such services must be able to perceive and make sense of human activities, and respond and learn from implicit and explicit feedback obtained through interaction with humans.

Unfortunately, human behavior is complex and unpredictable. Even within a given context, actions and reactions vary from one individual to another. Unobtrusive services must be able to adapt to individual variations and to the evolutions of preferences for a given individual. Adaptation is fundamental to unobtrusiveness, and thus socially acceptable, services.

Human activity is situation dependent and does not necessarily follow plans. Services within intelligent environments need information about the current situation to respond correctly without disrupting human activity. A strong contextual model is necessary for representing humans within the environment and their activities. The situations of this model can be decorated with system services to provide to the users

in the environment. The construction process for such a model and the associated services accommodate and assimilate human preferences. We have investigated the use of machine learning methods in order to provide a technology in which naturally expressed human reactions is used as feedback to adapt the behavior of context aware services.

This article motivates and introduces a general approach for learning situation models from user feedback. In the first part of this article, we describe a method to automatically acquire an initial model with minimal human intervention. We then propose a method to adapt an existing situation model with user feedback given on appropriate system services. The proposed approach has been implemented and evaluated with experiments in our augmented meeting environment.

A scenario

Bob is dreaming of a new intelligent home. The new home provides services to make Bob's life easier and more convenient. Bob hopes to reduce all that technical stuff that he needs to switch on/off, regulate, configure, etc. His ideal environment should provide entertainment and communication services with little or no configuration, adapting according to his preferences with a minimum of disruption and feedback. Bob should only need to indicate which service he wants and the system should adapt accordingly.

For instance, Bob enjoys jazz music when he is eating on the couch, but he does not want to be disturbed when eating with his girlfriend. Bob is willing to give feedback for learning to the system by giving specific voice commands in the environment or even, if needed, by clicking on services to provide on his PDA.

To satisfy Bob, the environment needs to be equipped with visual and acoustic sensors. For example, video cameras and a video tracking system and microphone array with speech detection and recognition can provide basic information about Bob's current location and activity. Of course, hand crafting of detection routines is not sufficient for Bob's dream as he wants the system to evolve, constantly adapting to his preferences. Thus a general model of the environment needs to be designed and then adapted according to Bob's remarks. The situation model has proved to be very useful for this task, being applied to various problems and domains [2].

A situation is a temporal state describing activities and relations of detected entities (persons) in the environment. Perceptual information from the different sensors in the environment are associated with situations. The different situations are connected within a network. A path in this network describes behavior in the scene. System services to be provided are associated with the different situations in the network.

Initial Situation Model

In order to build a situation model for Bob's home, we start with an initial model describing Bob's very basic behavior. This model can be seen as default configuration

of the system for the environment, being general and providing little detail. The picture below shows an example of such a default situation model for Bob's home.

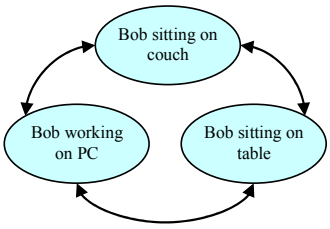

Fig. 1. Initial situation model for Bob's home

An open question, however, is from where to obtain a suitable initial model in order to minimize adaptation costs and user disruption. Normally, an expert constructs these models when setting up the system within environment. However, expert hours are expensive. An expert should at least be aided by automatic processes. In the following, we propose a schema for creating the initial situation network with a minimum of expert knowledge input. We exploit the addition of human expertise only for providing the situation labels. The derivations of the situations are, however, automatic, based on the recorded sensor signals.

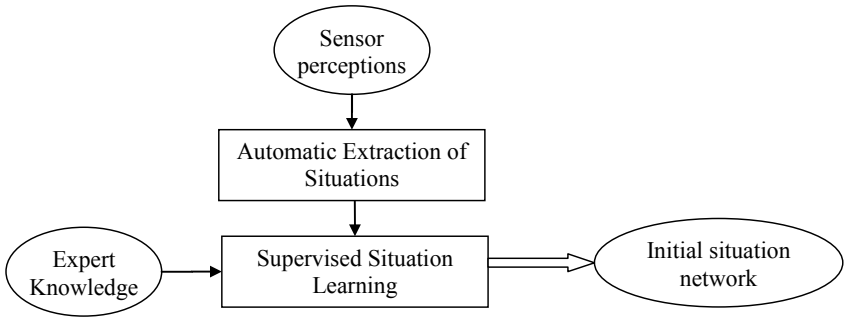

Fig. 2. Overview of the process of creating an initial situation model

The schema of the process of creating an initial situation model is depicted in Fig. 2. Given sensor recordings of Bob's activity in the environment, the automatic extraction of situations provides a first segmentation of sensor signals. This first segmentation is then labeled by an expert, integrating his expert knowledge. We assume here that expert knowledge and obtained automatic segmentation have a minimum of compatibility allowing the integration of expert situation labels. A supervised learning scheme is used to associate the situation labels with the recorded perceptions.

The method for initializing the situation model is based on the stream of perceptual events coming from different sensors in the environment. The method detects changes in the overall event distribution and provides a first segmentation based on the detected changes. The method has been tested on audio and audio-visual recordings of

small group and individual activities, producing an overall segmentation correctness of over 0.9 (Q value [9]) [3].

A supervised learning scheme [5] can be adopted to associate the situation labels with the recorded sensor perceptions (Fig. 3).

```
A. For each learner class do:
    a. {optimization step}
       For each situation label do:
         • Select learner/set of learners
         • Apply learner to given perceptions
    b. {validation step}
       Calculate ratio of between-situation and within-
       situation distance
    c. Repeat a.-b. until optimal ratio is obtained

B. Choose learner class with best ratio of between-
   situation and within-situation distance
```

Fig. 3. Scheme for supervised learning of situations

The idea is to use discriminant pattern recognition to recognize situations. Therefore, we first iterate over the possible learner classes (SVMs, Bayesian classifier, decision tree etc.) applicable for the task (part A). For each learner class, we select parameters and apply the learner to perceptual signals for each situation label (provided by an expert). The set of parameters that gives the best discrimination of the situations is retained (part A, step c.). Situations representations produced by the learner class with the best discrimination are retained (part B). As this is a general scheme, different forms of discriminant learning may be used interchangeably. The proposed supervised learning scheme has been evaluated on video surveillance data from the CAVIAR project, providing an overall recognition rate of 93.78 % [5].

The outputs of the supervised situation learning scheme are the situation representations for the initial situation network. The connections between the situations are constructed by considering the recorded sensor perceptions and existing transitions between the detected situations.

Integrating User Preferences

The initial situation model is simple, with insufficient detail about Bob's preferences. General situations, such as “Bob sitting on couch”, must be refined to obtain sub-situations incorporating the preferred system services. A possible adapted situation model could look like in Fig. 4.

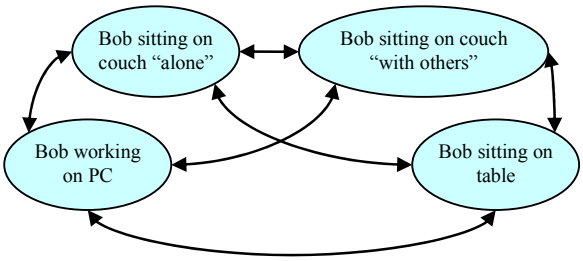

Fig. 4. Adapted situation model for Bob's home

"Sitting on couch" has been split into "Sitting on couch -alone-" and "Sitting on couch -with others-". Sensor perceptions need to be associated to the new sub-situations. Bob is, however, not interested in recruiting a system engineer implementing the new sub-situations. This refining process, and in particular the association of sensor perceptions to the new sub-situations, should hence be as automatic as possible. The new sub-situations need thus to be learned from recorded sensor perceptions as well as Bob's given feedback (via his voice or PDA) using machine learning methods. The supervised learning scheme (Fig. 3) can again be adopted to associate sensor perceptions to the new sub-situations.

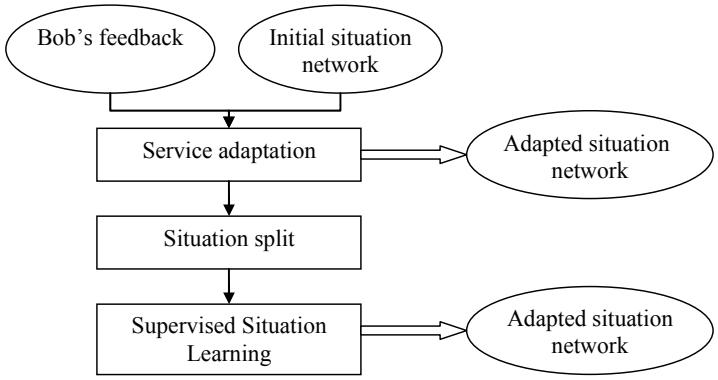

Fig. 5. Overview of the process of integrating user feedback into the initial situation model

The overall process of the automatic integration of Bob's feedback can be seen in Fig. 5. Bob's feedback and the initial situation network are the input. First, the algorithm tries to adapt simply the services associated to the given situations ("if service A is associated to S, and Bob indicates a preference for service B, then associate service B with S and dissociate service A"). If successful, the result is an adapted situation network integrating Bob's wishes. No situation split is necessary. However, if Bob's feedback indicates further that the concerned situation is too general, i.e. several distinct services are to be associated with one situation, the algorithm splits the situation into sub-situations. The sensor perceptions describing the new sub-situations

are learned using the supervised learning scheme. The proposed supervised scheme is run with Bob's feedback and the recorded sensor perceptions. Once the sub-situations are learned, they are inserted into the whole network by eliminating conflicts and erasing obsolete situations. The result is an adapted situation network with new sub-situations. A first evaluation in the PRIMA augmented office environment provided an overall performance of 94.3 % of correctly executed services for the evaluated scenarios [6].

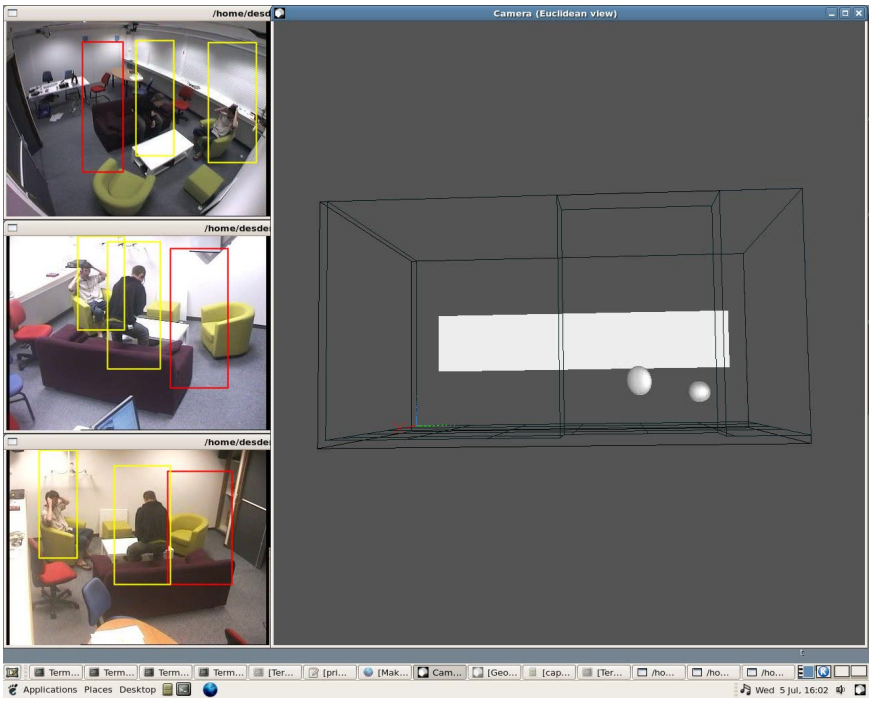

Fig. 6. 3D video tracking system fusing information of 3 2D trackers to a 3D representation

Experimental Evaluation

We evaluated the integral approach on 3 scenarios recorded in the PRIMA augmented home environment. The scenarios involved up to 2 persons doing different activities (introduction/address of welcome, presentation, aperitif, playing a game, siesta {1 person}) in the environment. A 3D video tracking systems [1] detected and tracked the people in the scene (Fig. 6.).
An individual activity detector [4] [8] has been applied to the extracted properties of the detected targets (persons). Additionally, the participants wore head microphones detecting whether the participant is speaking or not. A room microphone detected ambient noise in the scene. The association between video targets and head micro-

phones has been manually done by a supervisor at the beginning of each scenario recording. The observation generated by the system for each target is a code between 0 and 52, combining posture (sitting, standing, lying), movement (gesturing, changing position, walking), speech and ambient sound. The whole system runs in real-time, i.e. 25 frames per second.

The first step of our proposed approach is to create the initial situation model. We extract the situations from the sensor perceptions, i.e. the observations generated for the targets in the scene using our automatic segmentor [3]. The automatically extracted segments and the *ground truth* for the scenarios are depicted in Fig. 7. The overall segmentation exactitude *Q* [9] is best for scenario 2. This can be explained by the fact that the algorithm has difficulties to distinguish ground truth segments "game" and "aperitif". In scenario 1 and scenario 3, "game" and "aperitif" are detected as one segment. As in scenario 2 "playing game" and "aperitif" are separated by "presentation", these segments can be correctly detected.

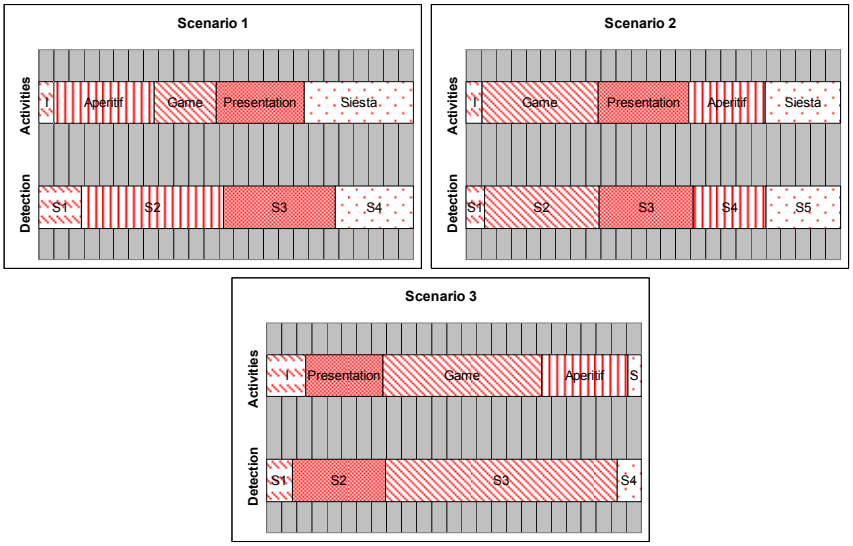

Fig. 7. Extracted situation segments and the corresponding *ground truth* for scenario 1 (Q = 0.68), scenario 2 (Q = 0.95), scenario 3 (Q = 0.74).

The supervised learning scheme is applied on the detected segments. As expert knowledge, we inject the situation labels: "introduction", "presentation", "group activity" (=aperitif or game), "siesta". We will adopt hidden Markov models [7] as unique learner class, iterating over left-right hidden Markov models of state numbers between 8 and 16 (=parameters of the class). To evaluate, we did 3-fold cross-validation, taking the detected segments + expert labels of 2 scenarios as input for learning and the third scenario as basis for recognition. As our system should be as responsive as possible, we evaluated different window sizes used for recognition. The results can be seen in Fig. 8. If we limit the observation time provided for recognition to 10 seconds (i.e. 250 frames with a frame rate of 25 frames/sec), we get a recogni-

tion rate of 88.58 % (Fig. 9). The recognition rate of "siesta" is poor due to the fact that in two of the three scenario recordings wrong targets have been created and detected when a person lay down on the couch, resulting in a disturbance of the existing target properties.

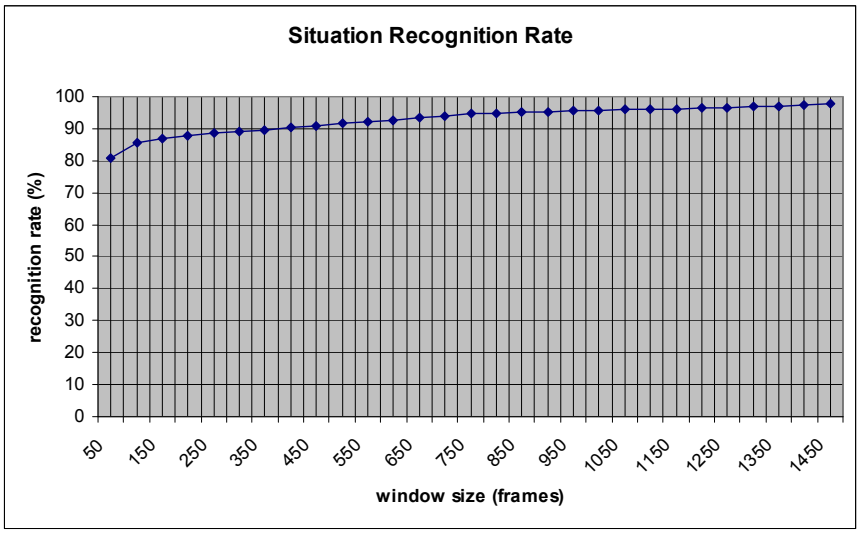

Fig. 8. Recognition rate of situations "introduction", "presentation", "group activity" (=aperitif or game) and "siesta" for different observation window sizes (in frames).

	introduction	group activity	presentation	siesta
introduction	0,97	0,00	0,00	0,03
group activity	0,00	1,00	0,00	0,00
presentation	0,03	0,03	0,83	0,11
siesta	0,40	0,00	0,08	0,52

	TP rate	FP rate	Precision	Recall	F-measure
introduction	0,97	0,07	0,41	0,97	0,58
group activity	1,00	0,02	0,98	1,00	0,99
presentation	0,83	0,02	0,95	0,83	0,89
siesta	0,52	0,03	0,72	0,52	0,61

Fig. 9. Confusion matrix and information retrieval statistics for each situation (observation window size=250 frames). The overall situation recognition rate is 88.58 %.

We have now learned an initial situation model with the situations "introduction", "group activity", "presentation" and "siesta". In order to integrate user preferences into this model, a user can give feedback to our system. The feedback is recorded and associated to the particular frame when it has been given. The initially learned model is then adapted according to this feedback (Fig. 5). For our scenarios, the user wants to integrate the following services:

- S1: Introduction → normal light and no music
- S2: Aperitif → dimmed light and jazz music
- S3: Game → normal light and pop music
- S4: Presentation → dimmed light and no music
- S5: Siesta → dimmed light and yoga music

The user gives one feedback indicating the corresponding service during each situation. As the initial situation model does not contain any situation-service associations, S1, S4 and S5 can then be simply associated to the corresponding situations. For S2 and S3, there is only one situation "group activity" which is too general in order to associate both distinct services. This situation needs thus to be split into sub-situations. The learned situation representation for "group activity" (here: a HMM) is erased and two distinct situation representations (here: HMMs) for "aperitif" and "game" are learned. The observations necessary to learn these situations are taken around the time points when the user gave the corresponding feedback. The size of the observation window used for learning the new sub-situations can be varied. We will adopt the double of the detection window size, i.e. 500 observation frames around the feedback time points to learn "aperitif" and "game". The obtained results of the 3-fold cross validation for recognition window size of 250 frames are depicted in Fig. 10.

	introduction	aperitif	game	presentation	siesta
introduction	0,97	0,00	0,00	0,00	0,03
aperitif	0,00	0,72	0,27	0,00	0,01
game	0,00	0,19	0,81	0,00	0,00
presentation	0,03	0,00	0,03	0,83	0,10
siesta	0,40	0,00	0,00	0,08	0,52

	TP rate	FP rate	Precision	Recall	F-measure
introduction	0,97	0,07	0,41	0,97	0,57
aperitif	0,72	0,08	0,70	0,72	0,71
game	0,81	0,09	0,81	0,81	0,81
presentation	0,83	0,02	0,95	0,83	0,89
siesta	0,52	0,04	0,72	0,52	0,60

Fig. 10. Confusion matrix and information retrieval statistics for each situation (observation window size=250 frames) after the split of "group activity". The window size for learning the new sub-situations is 500 frames. The overall situation recognition rate is 76.48 %.

Conclusion

In this paper we have described and evaluated a first approach for learning situation models in order to provide context-aware services. The obtained results are encouraging, but much remains to be done. First, the sensors necessary for a reliable

sensing of Bob's activities are not sufficiently reliable and require expensive installation. Multiple cameras, microphones or other sensors must be installed and calibrated in Bob's home. Clearly we will need automatic procedures for configuration and calibration of whatever sensors are used for such a system to be economically viable. Second, even though our results are encouraging, the error rates are still too high. Further improvements in detection and learning algorithms are necessary in order to provide a reliable system that could be accepted by Bob in his daily life. One way to alleviate this is to provide explanations to Bob. When errors occur (and corresponding system explanations are good), Bob could understand and correct wrong system perceptions and reasoning himself.

References

1. Biosca-Ferrer, A. and Lux, A., A visual service for distributed environments: a Bayesian 3D person tracker, submitted to International Conference on Computer Vision Systems 2007.
2. Brdiczka, O., Reignier, P., Crowley, J.L., Vaufreydaz, D., and Maisonnasse, J., Deterministic and probabilistic implementation of context. In Proceedings of 4th IEEE International Conference on Pervasive Computing and Communications (PerCom) Workshops, 2006.
3. Brdiczka, O., Maisonnasse, J., Reignier, P., and Crowley, J.L, Extracting activities from multimodal observation. In Proceedings of 10th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES), 2006.
4. Brdiczka, O., Maisonnasse, J., Reignier, P., and Crowley, J.L., Learning individual roles from video in a smart home. In 2nd IET International Conference on Intelligent Environments, Athens, Greece, July 2006.
5. Brdiczka, O., Yuen, P.C., Zaidenberg, S., Reignier, P., and Crowley, J.L., Automatic acquisition of context models and its application to video surveillance. In Proceedings of 18th International Conference on Pattern Recognition (ICPR), 2006.
6. Crowley, J.L., Brdiczka, O., and Reignier, P., Learning situation models for understanding activity. In Proceedings of 5th International Conference on Development and Learning (ICDL), 2006.
7. Rabiner, L. A., A tutorial on Hidden Markov Models and selected applications in speech recognition. In Proceedings of IEEE 77(2):257-286, 1987.
8. Ribeiro, P., Santos-Victor, J., Human activity recognition from Video: modeling, feature selection and classification architecture. In Proceedings of International Workshop on Human Activity Recognition and Modeling, 2005.
9. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., and Lathoud, G., Multimodal Group Action Clustering in Meetings. In Proceedings of International Workshop on Video Surveillance & Sensor Networks, 2004.