Ambient Intelligence through Image Retrieval

Jean-Marc Seigneur, Daniel Solis*, Fergal Shevlin

Department of Computer Science, Trinity College Dublin,
* and Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya
Jean-Marc.Seigneur@trustcomp.org
solisand@tcd.ie
Fergal.Shevlin@cs.tcd.ie

Abstract. An ambient intelligent environment needs dynamic enrollment of strangers without too much human intervention. For this purpose, we propose an entity recognition process based on images captured with low-cost but widespread webcams and easy-to-deploy image processing techniques. We find that the use of levels of confidence in recognition due to different techniques and context-based image retrieval improves the process.

1 Introduction to Ambient Intelligence

Weiser's vision of ubiquitous computing [14] will be realised when computing capabilities are woven into the fabric of everyday life. Already companies in the domestic appliance market are promoting their smart home appliances – with communication, computation and data storage capabilities. Through such initiatives, spaces will become smart: endowed with ambient intelligence (AmI) to enhance the user's experience. However, challenges remain for the fulfilment of this vision. These include auto-configuration and autonomy, especially with respect to security [8]. The solution may come from a "concierge" process aware of what happens in the space, which can recognise strangers, acquaintances, friends or foes.

Vision is an obvious mechanism for the recognition of people in spaces. It has been used for authentication based on visual biometrics (such as fingerprint, face or gait recognition [1, 6, 10]). Generally, these techniques are used in controlled environments, where enrollment is mandatory (i.e., persons to be enrolled have their visual biometrics entered into the security system in advance). This contrasts with the fundamental requirement for ubiquitous computing environments: to allow for potential interaction with unknown entities [9]. In an AmI environment, enrollment cannot always require human intervention, e.g., from a system administrator. A smart space is not an improvement if it makes busy householders even busier. In public environments, there is no list of known people to be enrolled. People roam from one space to another as they wish. This introduces the requirement for smooth dynamic enrollment, i.e., the door should not be closed to strangers, but instead any stranger presenting themselves might become an acquaintance. To allow for dynamic enrollment of strangers and unknown entities, we have proposed an entity recognition (ER) process [9].

In this paper we investigate image retrieval as part of an ER scheme, called the vision entity recognition scheme (VER). We use commercial-off-the-shelf (COTS) products (e.g., basic "webcam" shipped with PC) in order to get an idea of what could be ubiquitously achievable today. The ER process allows the use of any scheme (i.e., even weak or unreliable) by taking into account confidence in recognition. We investigate how to improve indexing and retrieval of previously recorded imagery based on its context (e.g., time and weekday) in addition to its content.

2 Applying Vision Techniques to ER: VER

One of the foundations of security is authentication. Stajano [12] emphasized that without being sure with whom an entity interacts, the three fundamental properties – confidentiality, integrity and availability – can be trivially violated. Usually, authentication is the first step to ensure security in computing environments but other work [3] discusses why traditional authentication should be reconsidered for pervasive computing. Our end-to-end trust model [9] addresses this problem by starting with recognition, which is a more general concept than authentication, i.e., entity recognition encompasses authentication. To allow for dynamic enrollment of strangers and unknown entities, we have proposed the entity recognition (ER) process, which consists of four steps:

1. Triggering of the recognition mechanism
2. Detective Work to recognize the entity using the available recognition scheme(s)
3. Discriminative Retention of information relevant for possible recall or recognition
4. Upper-level Action based on the outcome of recognition, which includes a level of confidence in recognition

As an example of what is possible with this approach, we have developed an entity recognition component based on pluggable recognition modules (PRM), which allows the integration of more or less secure recognition schemes. We conjecture that the ability to recognise another entity, possibly using any of its observable attributes, is sufficient to establish trust in that entity based on past experience. The “Resurrecting Duckling” security policy model [13] is an example of entity recognition; ducklings know that their mother is the first entity who sent the imprinting key when they were born. They must be able to recognise when the entity with which they interact is the one who sent the imprinting key, no more. Most of the ER schemes cannot be considered as strong recognition schemes. However, we can still use them in our recognition process since the outcome of the ER process provides meta-data on the level of confidence in recognition including technical trust in the recognition scheme used.

In our current prototype, we assume a room with one door (see Fig. 2) and the equipment described in the Appendix. We combine different image processing and retrieval techniques (discussed in Section 3 and Section 4) to recognise people entering and leaving the room. The ER process allows recognition of previously observed/encountered entities based on visual evidence, i.e., imagery. There is a PRM where different vision schemes can be implemented (e.g., face matching or clothes colour). Each time someone moves in front of the camera, the ER process (depicted in Fig. 1) is triggered: we call this self-triggering because the system itself takes the initiative to start the recognition process in order to recognize potential surrounding entities. In step 2 of the ER process, the detective work consists of carrying out a variety of visual analyses to obtain a level of confidence of each recognition. Retrieval of previous imagery is based on content as well as context (see Section 4).

Step 3 is closely related to step 2 because discriminative retention of recognition must be based on previously stored imagery. A difficult question is to define when the person who enters the room is new and converge to the real number of different persons monitored so far. In the ER process, there is no initial mandatory enrollment but enrollment is moved down in the process and occurs at step 3 when recognition information on a new entity is stored for the first time and for later recall. A person is digitally represented by a principal (*p*). The indexing of stored imagery for future retrieval at the end of step 3 also makes use of context (see Section 4). Step 4 of the ER process concerns further actions to be taken according to what person is recognised. This is almost beyond the scope of the paper but it may also be used to increase the level of work to be done during one round of ER. For example, if a new person is recognised at 2am, the concierge should increase its level of suspicion (and maybe send a warning message to the security guards) as well as increase the level of detective work and discriminative retention (which may augment the chance to later recognise the potential theft).

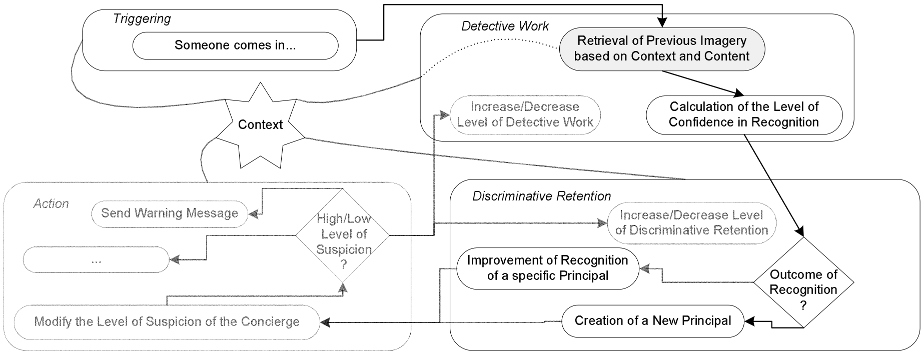

Triggering
Someone comes in...
Detective Work
Retrieval of Previous Imagery based on Context and Content
Increase/Decrease Level of Detective Work
Calculation of the Level of Confidence in Recognition
Context
Action
Send Warning Message
High/Low Level of Suspicion ?
...
Modify the Level of Suspicion of the Concierge
Discriminative Retention
Increase/Decrease Level of Discriminative Retention
Improvement of Recognition of a specific Principal
Outcome of Recognition ?
Creation of a New Principal

Fig. 1. VER process diagram

3 Image Processing Techniques

Due to the important requirement that the system needs as little as possible set-up or calibration by the owner of the space, the techniques used for image segmentation and analysis are necessarily simple. Additionally, the near-real time performance requirements of the system preclude the analysis of complex biometric characteristics such as gait, but we have designed our indexing and retrieval scheme to allow the inclusion of such characteristics should sufficient computational power exist.

3.1 Feature Segmentation

Simple inter-frame image subtraction allows motion to be identified. If the motion blob area exceeds a certain threshold then it is considered a potential person. The region-merging via boundary-melting algorithm [11] is applied to segment the blob into distinct regions of significance for recognition. The significant regions, as shown in Fig. 2, are:

a) Skin. Using the approach to skin segmentation suggested by Perez et al. [7], we transform from (R,G,B) colour space into the normalised (RN,GN) model and classify a pixel as skin if its values lie between certain upper and lower thresholds in RN and GN.

b) Face. The uppermost region of skin exceeding a certain area threshold and with appropriate elongation is considered to be the face. Its bounding rectangular region extracted. If the region is larger or smaller than 40x40 pixels then it is sub- or super-sampled as appropriate to facilitate inter-image comparison.

c) Clothes. Non-skin regions exceeding a certain area are considered clothes. There are typically two such regions found: top and bottom.

d) Hair. In theory it should be relatively straightforward to segment hair, using its colour as another feature to facilitate recognition. However in our environment was insufficient contrast between the hair and the background for it to be segmented reliably.

e) Height. Relative height can be approximated as the difference between the highest and lowest segmented pixel. Any height comparison must take into account the position of the feet.

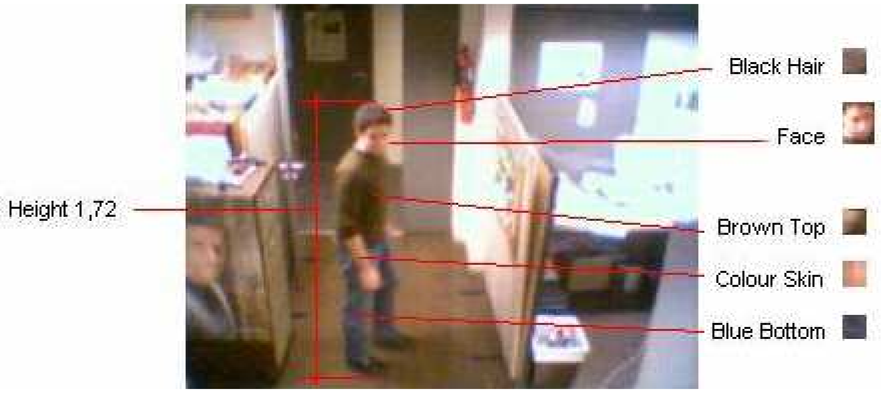


Fig. 2. Environment and segmented features

3.2 Feature Analysis

The face is the only feature that can be used for recognition with any reasonably high degree of confidence (as shown in Subsection 5.2). To date we have used simple template-matching (normalised cross-correlation) to match segmented faces. However a principal components analysis (PCA) approach could also be used and would probably yield better results, as has been shown in [5].

4 Context-based Image Retrieval

Each time a face is segmented from the real-time video sequence, it is appended to a list. When the sequence is finished, each face of the list is compared to the set of different segmented faces stored previously. If there is no match above a minimal level of confidence, or no faces have been stored previously, it is added to the list of observed faces. Details of the other segmented features (for example, clothes colour) are associated with the face, as are temporal attributes such as date, time, and day of week. If a face matches above the minimal level of confidence then the other details are retrieved and used in the recognition process. In our approach, there is no training data and database of known users per se due to the requirement of dynamic enrollment. This differs from related work on real-time vision-based multi-modal recognition [10].

The advantage of pervasive computing environments is that computing entities are context-aware – environmental information that is part of an application's operating environment can be sensed by the application. Castro and Muntz [2] pioneered the use of context for multimedia object retrieval. We apply the concept within our ER process, which enables the concierge to adapt retrieval and recognition based on context and level of suspicion without the help of an administrator. Crowley et al.'s software architecture [4] for observing and modelling human activity built on top of their ontology for context-awareness is valuable for our type of application scenario. In their approach, our concierge may be seen as a supervisory controller of the ER process, which corresponds to an entity grouping of observational processes. Dey defines context as "any information that can be used to characterize situation" [4]. We especially make use of time and date to index and retrieve imagery.

4.1 For Indexing

The first time the VER scheme is started in a new space, the list of faces and associated visual and temporal attributes is empty. As soon as someone comes in front of the camera, a sequence of faces is extracted from the video. Associated with each sequence is a structure storing the other elements of specific context. Our proof-of-concept implementation consists of storing the time and the day of the week. For each sequence, height and colour information is also computed. A database is used to store the recognition clues extracted from each sequence. These recognition clues are indexed in specific rows and each row consists of a supposed different person. We can then dynamically index the different rows based on context similarity. For example, we can order the rows decreasingly from the row which contains images the most often seen on Monday mornings around 8am. For performance reasons, each sequence of images is processed for face template matching after the end of the sequence when nobody is moving in front of the camera. The face template matching process is too expensive to be run in parallel during the capture of the sequence.

Once all the images of the sequence are compared, we obtain a probability distribution of the following form:

$$(N_{PFR1}+N_{FR1})*p_1+\ldots+(N_{PFRi}+N_{FRi})*p_i+\ldots+(N_{PFRn}+N_{FRn})*\, p_n+N_{unknown}*p_{unknown}+N_{discarded}*p_{discarded}$$

where p_i is the supposed different person i among n previously seen persons, N_{PFRi} is the number of perfect face recognitions (match above PerfectFaceRecognition) of person i, N_{FRi} is the number of face recognitions (match below PerfectFaceRecognition but above UnknowFaceRecognition) for the person i, $N_{unknown}$ is the number of faces either of a new person or a very different face profile of a known person (match below UnknowFaceRecognition but above BogusFaceDiscarded) and $N_{discarded}$ is the number of images considered to be of bad quality (below BogusFaceDiscarded). From this distribution, a choice has to be made. Is it a new person or should it update the recognition clues of a previously known person? The update only consists of faces that are considered different enough that previous images (that is, between PerfectFaceRecognition and BogusFaceDiscarded) in order to improve scalability. In cases where face recognition confidence is borderline, we use the other visual attributes to help in the decision-making process. So far, we have followed an empirical solution, which has given encouraging results in real settings. However, we have chosen to discard sequences which might pollute the database with poor quality face images. The simplified pseudo-code of the algorithm is depicted below:

```
Pick the person i with the greatest ($N_{PFRi}$+$N_{FRi}$)
if($N_{PFRi}$>(10%*TotalOfNotDiscardedImages))
        UpdateFacesOfPerson$_i$
if(NoPerfectMatch)
{
        if($N_{FRi}$>50%*$N_{unknown}$)
                if (((HeightMatching*50%)+(ColourMatching*50%)) >= 50%)
                                UpdateFacesOfPerson$_i$
        if($N_{unknown}$>50%*$N_{FRi}$)
                if (((HeightMatching*50%)+(ColourMatching*50%)) < 50%)
                                CreateNewPerson
}
```

4.2 For Retrieval

Thanks to our indexing, we can prioritize the retrieval based on context (i.e., time and day of the week). There are four parameters used in our algorithm: TimeAndDayOfWeeK; PerfectFaceRecognition (i.e., the percentage threshold above which the recognition match is considered perfect: empirically from the reading of several sequence processing samples say 92%), UnknowFaceRecognition (that is, the percentage threshold below which the recognition match is considered either a new person or a very different face profile of a known person: again empirically say 85%) and BogusFaceDiscarded (i.e., the percentage threshold below which the recognition match indicates that the image does not correspond to a face and is discarded: we empirically chose 30%). In order to benefit from a probabilistic approach and the fact that the images of a same sequence correspond to the same person (who is entering the room), at most 30 faces are extracted from the sequence and compared to all previous faces stored in the database. In order to speed up the process, the comparison is stopped if PerfectFaceRecognition is reached and the images stored in the database are reordered. The reordering consists of presenting the images of the previously recognised person first, ordered by their number of previous matches. Section 5.1 evaluates this retrieval approach.

5 Evaluation of the Retrieval System

We start by comparing random-based and context-enhanced retrieval and indexing. Then, we discuss which vision techniques are more important for improving the accuracy and relevance feedback of retrieval.

5.1 Context vs. Random for Retrieval and Indexing

One of the reasons we chose a context-based retrieval and indexing rather than retrieval with a random order of images is to obtain a faster retrieval scheme. It is worth mentioning than stopping the retrieval and not assessing all stored images for each new image is faster but we lose the opportunity to detect a recognition result greater than the PerfectFaceRecognition. However, this allows us to compare if context-based is really faster than random-based retrieval.

For this assessment, videos of 10 different persons (including Europeans, Chinese and Indians) entering the room 4 to 5 times were recorded. The database was populated with the same sequence for each person: these 10 sequences resulted in 205 faces stored in the database. Then, the remaining sequences of each person were processed (although no update/creation was applied) and resulted in the extraction of 757 faces. Using a random approach, this corresponds to 155185 matches (757*205). Thanks to our PerfectFaceRecognition bound and context reordering (explained in Subsection 4.2), assuming that each match takes the same time, the process was roughly 1.4 faster (that is, 44780 fewer matches were needed).

5.2 Empirical Assessment of the Technical Trust of Each Vision Technique

Each recognition scheme has to be assessed concerning its technical trustworthiness, which can be seen as the relevance feedback of retrieval. The number of people we used for this assessment (i.e., 10) is in line with the assessment done in previous related work [10] (i.e., 12).

The outcome of the ER process can be a set of *n* principals (*p*) associated with a level of confidence in recognition (*lcr*). The VER scheme is proactive: it triggers itself and uses a range of vision techniques which give evidence to compute a probability distribution of recognised entities. For example, a person among *n* persons previously recognised enters a room which is equipped with a biometric ER scheme. The outcome of recognition demonstrates hesitation between two persons: p_2 and p_3 are recognized at 45% and 55% respectively, these percentages represent the level of confidence in recognition. So, all other principals are given a *lcr* of 0%. We have:

$$OutcomeOfRecognition = \sum_{i=1}^{n} lcr_i\, p_i = 0*p_1+0{,}45*p_2+0{,}55*p_3+\ldots+0*p_i+\ldots+0*p_n$$

Technical trust is associated with each vision technique (*vt*): for example face template matching is *vt1* with *tt1*. Each technique provides a level of recognition (*lr*) for each principal. Assuming that we have *m* vision techniques and that each technique is weighted (with *w*) comparatively to other ER schemes used, we have:

$$\mathrm{lcr} = \sum_{j=1}^{m} lr_j * tt_j * w_j$$

If the sum of lr is too low, this suggests that we need to create a new principal.

Practically, for each different vision recognition technique (face, height and colour), we populated the database as in the previous subsection with 10 persons and then for each remaining sequences (3 or 4 different sequences for each person and 36 sequences in total), we counted how many times each scheme makes the right decision (that is, if the sequence corresponds to person *i*, the scheme should recognise person *i*). We obtained a technical trust of: 0.94 for face template matching (34/36), 0.39 for height matching (14/38) and 0.53 for colour matching (19/36).

6 Conclusion

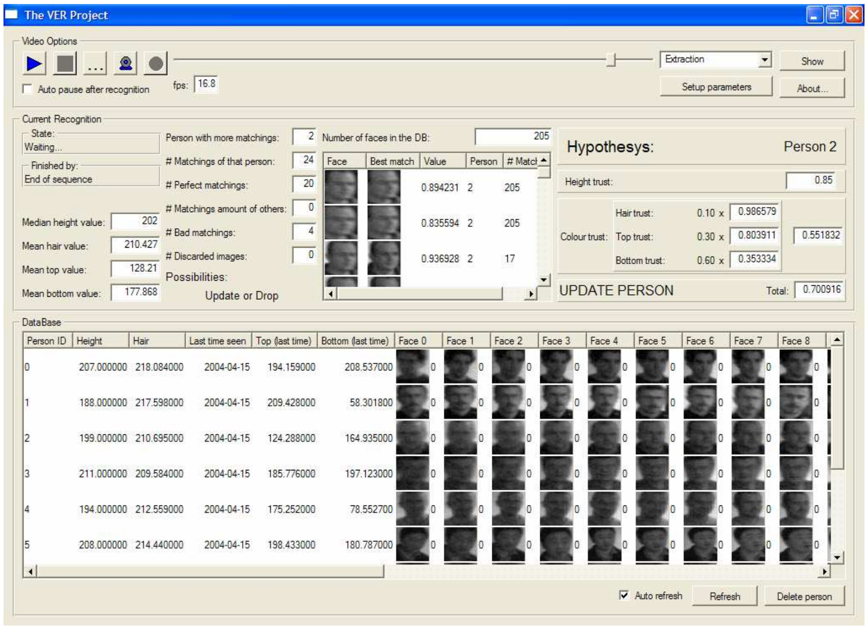
This work demonstrates the applicability of our entity recognition process to computer vision techniques. The use of context and level of confidence in recognition allows us to index and retrieve faster than with a random approach. In fact, the convergence to the true number of people is based not only on image content but also on context. We obtain dynamic enrollment. However, in order to get a database converging to the real number of persons, the system still drops lots of sequences, which are considered of bad quality and could pollute the database.

We have determined that low-cost webcam camera imagery and simple image processing and analysis techniques are sufficient to allow face recognition with a reasonable level of confidence, and other visual attribute recognition with lower, but still useful, levels of confidence. Initial evaluations of the system have yielded promising results and shown that little configuration is needed.

We argue that what is achieved by our system can already be useful for a range of ambient intelligent applications, especially for applications focusing on monitoring rather than security. We speculate that the widespread deployment of our system (which can be done since webcams are widespread and if we release our software) could already raise serious privacy concerns.

7 Appendix: Equipment and Software

The COTS low-cost CCD camera with USB interface to a conventional laptop (PentiumIII mobile CPU 866MHz with 256MB RAM) is typical of what is often referred to as a "webcam". It is used in a mode which provides 320x240 pixel 8 bit colour imagery at 15Hz for each channel. Actual resolution and sensitivity are lower due to a colour filter over the CCD and the poor-quality analog-to-digital converter used for quantisation. The camera's focal length is 30mm. Its lens faces the door. The software is written in C++ and uses a MySQL database. The graphical user interface (GUI) is presented in Fig. 3.



□ Auto pause after recognition

fps: 16.8

Extraction

State: Waiting...

Finished by: End of sequence

Median height value: 202

Mean hair value: 210.427

Mean top value: 128.21

Mean bottom value: 177.868

Person with more matchings: 2

Matchings of that person: 24

Perfect matchings: 20

Matchings amount of others: 0

Bad matchings: 4

Discarded images: 0

Possibilities:

Update or Drop

Number of faces in the DB: 205

Face	Best match	Value	Person	# Match
		0.894231	2	205
		0.835594	2	205
		0.936928	2	17

Hypothesys: Person 2

Height trust: 0.85

Colour trust:

Hair trust: 0.10 x 0.986579

Top trust: 0.30 x 0.803911

Bottom trust: 0.60 x 0.353334

0.551832

UPDATE PERSON

Total: 0.700916

Person ID	Height	Hair	Last time seen	Top (last time)	Bottom (last time)	Face 0	Face 1	Face 2	Face 3	Face 4	Face 5	Face 6	Face 7	Face 8
0	207.000000	218.084000	2004-04-15	194.159000	208.537000	0	0	0	0	0	0	0	0	0
1	188.000000	217.598000	2004-04-15	209.428000	58.301800	0	0	0	0	0	0	0	0	0
2	199.000000	210.695000	2004-04-15	124.288000	164.935000	0	0	0	0	0	0	0	0	0
3	211.000000	209.584000	2004-04-15	185.776000	197.123000	0	0	0	0	0	0	0	0	0
4	194.000000	212.559000	2004-04-15	175.252000	78.552700	0	0	0	0	0	0	0	0	0
5	208.000000	214.440000	2004-04-15	198.433000	180.787000	0	0	0	0	0	0	0	0	0

☑ Auto refresh

Fig. 3. Software GUI

8 References

[1] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and G.-R. J., "Multimodal Biometric Authentication using Quality Signals in Mobile Communications", in *Proceedings of the 12th International Conference on Image Analysis and Processing*, IEEE, 2003.

[2] P. Castro and R. Muntz, "Using Context to Assist in Multimedia Object Retrieval", in *ACM Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999, http://www.info.uqam.ca/~misrm/papers/castro.ps.

[3] S. Creese, M. Goldsmith, B. Roscoe, and I. Zakiuddin, "Authentication for Pervasive Computing", in *Proceedings of the First International Conference on Security in Pervasive Computing*, 2003.

[4] J. L. Crowley, J. Coutaz, G. Rey, and P. Reignier, "Perceptual Components for Context Aware Computing", in *Proceedings of Ubicomp'02*, 2002, http://citeseer.nj.nec.com/541415.html.

[5] C. Czirjek, N. O'Connor, S. Marlow, and N. Murphy, "Face Detection and Clustering for Video Indexing Applications", in *Proceedings of Advanced Concepts for Intelligent Vision Systems*, 2003, http://www.cdvp.dcu.ie/Papers/ACIVS2003.pdf.

[6] A. K. Jain, A. Ross, and S. Prabhakar, "An Introduction to Biometric Recognition", in *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 2003.

[7] C. Perez, M. A. Vicente, C. Fernandez, et al., "Aplicacion de los diferentes espacios de color para deteccion y seguimiento de caras.", in *Proceedings of XXIV Jornados de Automatica*, Universidad Miguel Hernandez, 2003.

[8] J.-M. Seigneur, C. Damsgaard Jensen, S. Farrell, E. Gray, and Y. Chen, "Towards Security Auto-configuration for Smart Appliances", in *Proceedings of the Smart Objects Conference*, 2003, http://www.grenoble-soc.com/proceedings03/Pdf/45-Seigneur.pdf.

[9] J.-M. Seigneur, S. Farrell, C. D. Jensen, E. Gray, and Y. Chen, "End-to-end Trust Starts with Recognition", in *Proceedings of the Conference on Security in Pervasive Computing*, LNCS 2802, Springer, 2003.

[10] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated Face and Gait Recognition From Multiple Views", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[11] M. Sonka, V. Hlavac, and R. Boyle, "Image Processing, Analysis, and Machine Vision", Second Edition, PWS Publishing, 1999.

[12] F. Stajano, "Security for Ubiquitous Computing", ISBN 0470844930, John Wiley & Sons, 2002.

[13] F. Stajano and R. Anderson, "The Resurrecting Duckling: Security Issues for Ad-hoc Wireless Networks", in *Proceedings of the International Security Protocols Workshop*, 1999, http://citeseer.nj.nec.com/stajano99resurrecting.html.

[14] M. Weiser, "The Computer for the 21st Century", Scientific American, 1991, http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html.