M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 500
submitted to ISWC'99

Attentional Objects for Visual Context Understanding

Bernt Schiele and Alex Pentland

The Media Laboratory, Room E15-384a,
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139
email: {bernt,sandy}@media.mit.edu

Abstract

This paper exploits wearable computers' unique opportunity to record and index the visual environment of the user from the "first-person" perspective. We propose to use a hat-mounted wearable camera to record what the user sees during the day with a wearable computer. This camera can be used to make the computer more contextually aware of the user and their actions. Furthermore, the camera can be used to record, analyze and index the visual environment of the user. By keeping track of the actions of the user upon and within the environment the system can be more aware of the interactions of the user within the environment. An important aspect of the system is to automatically extract objects of user interest, and their motion within the environment and relative to the user.

1 Introduction

Wearable computers have the potential to "see" as the user sees, "hear" as the user hears, and experience the life and the environment of the user in a "first-person" sense. As has been pointed out in [23], wearable computers offer a unique opportunity to recover more general user context. The importance of context for communication and interface can not be overstated. Without context awareness, wearable computers will only be useful for particular purposes or even more to the point, will be only useful in a particular context. See also [10] and [16].

Sensors, such as a head-mounted camera, not only allow modeling and recognizing the user's context but can be used to record and analyze the user's environment from a "first-person" perspective. This paper proposes a wearable system which records the visual field of view continuously onto the user's wearable computer while exploiting the habits of humans (such as fixating objects of interest) to structure and index this visual record.

The system is designed to model, recognize and track objects and people automatically in the environment of the user. The system can structure the visual environment based on these objects and persons. By observing the absolute and relative motions of objects and people the system can reason about the interactions of people and objects within the environment. The duration for which a person fixates a moving object is a powerful indicator for the potential interest of that object or that person. The physical interaction of the user with an object (e.g. by touching or picking up) is another indicator of the relevance of these objects to the user. Such *attentional objects* and people can be used to analyze, index and structure the video-database stored on the wearable computer.

The fact that the camera is hat-mounted and therefore reflects a first-person view of the environment enables the exploitation of the habits of humans such as fixating objects of interest. The system also can use the relative motion of objects with respect to the user. The ability to use the special nature of a wearable computer to extract objects of interest makes this system very different to the standard systems for video-browsing and -indexing. By exploiting the habits of the user we can extract structure from the video stream which reflects the relevance and importance of various aspects of the visual environment to the user and is therefore useful to the user.

Our aim is to build a wearable system which is aware of the user context and continuously models and recognizes the environment of the user as well as the user's interests and interactions with the environment. In order to put the remainder of the paper into perspective, the next section summarizes several systems which have been designed to be aware of different aspects of the user's context. Section 3 discusses the usefulness and the technical requirements for a wearable system which continuously observes the visual environment of the user. Section 4 describes a system which uses a state-of-the-art object recognition module to recognize previously recorded objects in the visual field using a wearable camera. Section 5 discusses in more detail a

system which automatically extracts and tracks objects in the visual environment.

1.1 Related Work

This section briefly describes several systems which are related to the system proposed here.

Memory Augmentation: Memory augmentation has evolved from simple pencil and paper paradigms to sophisticated personal digital assistants (PDAs) and beyond. Some related memory augmentation systems include the "Forget-me not" system [13], which is a personal information manager inspired by Weiser's ubiquitous computing paradigm, and the Remembrance Agent [20], which is a text-based context-driven wearable augmented reality memory system. Both systems collect and organize data that is relevant to the human user for subsequent retrieval.

Augmented Reality: Augmented reality systems form a more natural interface between user and machine. In [11] a virtually documented environment system is described which assists the user in some performance task. It registers synthetic multimedia data acquired using a head-mounted video camera. However, information is retrieved explicitly by the user via speech commands. [19] uses contextual tags (such as infrared and visual markers) in order to enhance and the context of the environment. [8] presents a system which displays interesting pre-stored information about the environment of the user. Such a system can be used as a "touring machine".

Personal Imaging: Mann [14] proposes to construct high-resolution images from the visual environment using a head-mounted wearable camera. Essentially, the system can construct a high-resolution image out of many smaller overlapping images. The resulting image can be seen as a photographic memory of the visual environment of the user.

Video Indexing and Retrieval: Many system exist which attempt to classify, index, browse and retrieve videos from databases. See for example [25, 17, 18, 4]. Most systems however, only use global features such as color histograms for each frame of the video in order to segment the video-stream and to index the video. An notable exception is a system proposed by Chang et al. [5] where individual regions are extracted from the video frames. These regions are tracked over time enabling queries to the video database such as: find region of that color moving in this particular motion pattern. Using such a database query, the authors have been able to retrieve the video-clip of a high-jumper. They note the importance of extracting and matching of the motion of regions in order to index the video database.

Remote Sensing Wearable computers allow collaboration as well as remote sensing of distant environments. [3, 12] present a wearable system with remote sensing capabilities in order to allow cooperation of mobile technicians.

2 Context Aware Perceptual Systems

User context includes such things as time of day, location of the user and user's action. In the following we briefly describe several systems which have been designed to recognize the user's location without using infrastructure such as RF tags or landmarks. We also describe two systems which are using a user-observing camera to recognize the user's gestures and actions.

2.1 Recognizing the user's location

Three different systems have been developed in our lab to recognize the location by means of a wearable computer. The first system [23] uses two head mounted wide-angle cameras (see figure 1). The first camera is pointed forward whereas the second is pointed downwards, enabling the observation of the floor. The system continuously calculates the color mean values of three image patches. For each room of the floor of an MIT building, an HMM has been trained to model and recognize each individual room. Since every room can only be reached by a small number of doors, we can train and employ a statistical grammar for the floor which allows to recognize the user's location with an accuracy of about 82%.

Figure 1: Above: The two camera hat. Below: the downward- and forward-looking camera views.

The second system [2] also uses a head mounted camera in order to recognize the user's location. Here, the

system not only recognizes a particular location but also stores information about the approaching trajectory to the location. By using dynamic programming to compare image sequences of different approaches to particular locations, the system can match the incoming video in real-time with a dictionary of locations of the environment. Whereas the first system continuously recognizes the location and relies on the statistical grammar in order to restrict the search space, this system can be used to recognize particular locations in the environment for recovery and reentry into a particular environment.

The third system, the Environmentally-A-Wearable (EW) [6] uses auditory and visual cues to classify the user's environmental context. Like "the fly on the wall" (except now the fly is on your shoulder) it does not try to understand in detail every event that happens around the user. Instead, EW makes general evaluations of the auditory and visual ambiance and whether a particular environment is different or similar to an environment that the user was previously in. In order to make use of the audio-visual channel detectors for specific events are constructed. Events can be simple such as a bright light and loud sounds, or more complicated such as speaker sounds and objects. Given a set of detectors higher order patterns can be observed. For example, a user's audio-visual environment can be broken into scenes (possibly overlapping) such as 'talking to a person', 'visiting the grocery store', 'walking down a busy street', or 'at the office' that are collections of specific events such as 'footsteps', 'car horns', 'cross-walks', and 'speech'. We can recognize scenes by using detectors for low-level events that make up these scenes. This identifies a natural hierarchy in a person's audio-visual environment. For further details see [6].

2.2 User-observing wearable cameras

By using a head-mounted camera pointing down, the user himself can be tracked (see figure 1). This novel viewpoint allows the user's hands, feet, torso, and even lips to be observed without the gloves or body suits associated with virtual reality gear. The hat-mount of Figure 1 provides a surprisingly stable mounting point for the camera. Different lighting conditions can be normalized for by using the constantly visible nose as calibration object.

The user-observing wearable camera has been used to observe and detect the user's hand gestures in the context of American sign language recognition [24]. Another system [23] uses the user-observing camera to recognize the current task of the user. These systems demonstrate a set of tools directed at recovering user context. Specifically, complex sets of time varying signals (i.e., gestures and user actions) from a self-observing body-mounted camera through the use of color blob analysis and HMM's.

3 Environment observing wearable camera

As we have pointed out earlier our goal is to build a system which is not only aware of the context of the user but also analyzes and reasons about the visual environment of the user. The reminder of the paper therefore exploits the possibilities for a wearable computer which uses a hat-mounted camera in order to observe the visual environment from a "first-person" perspective. This section discusses potential applications and the technical requirements for a system which continuously records and analyzes the visual environment of the user. Section 4 describes a wearable system which continuously recognizes objects in the users' visual field. Section 5 discusses a system to automatically extract objects and object parts from the visual environment which are of potential interest to the user.

In order to emphasize the usefulness of the analysis of the visual environment from a "first-person" perspective we want to give a number of examples:

- During a conversation we are typically looking at each others face, over extended periods of time. More generally humans have the tendency to fixate an object of interest over extended periods of time such as a piece of art or a street-map. Since these objects of interest are present in the "first-person" view of the system for some time, the system is able to extract a model of that object of interest. Such a model can be used to index into and structure the stored video-database and to recognize the object again later.

- People's gaze often follows (or literally tracks) objects of interest. Such a tracked object may be another person or a car. By using the hat-mounted camera, the tracked object is relatively still with respect of the rest of the scene, which is moving. Therefore, by analyzing the relative motion of "things" in the visual environment, the system could segment objects. Furthermore, since the objects are tracked by the user for a certain amount of time, the system can also hypothesize the respective relevance of these objects to the user.

- Several systems have been developed to continuously track human hands (see for example [24]). Knowing the position of the hands in the environment, the system can extract objects from the environment which the user touched, moved or otherwise handled. As before, the system can use this

information in order to estimate the respective importance of the objects to the user. Eventually, the system might even be able to help the user to find misplaced objects.

All the above examples exploit the fact that the hat-mounted camera has approximately the same field of view as the user. What is required for the above examples is the *extraction of regions* from the visual environment of the user which might correspond to entire objects or objects parts. By modeling and tracking these regions over time we can analyze their respective motion relative to the user and relative to the environment. This information is extremely useful to hypothesize the respective importance of objects to the user, as the above examples show. Furthermore, by modeling the user's action upon and in the environment the system might model the interactions of the user with the environment. This which can be of great help to a wearable system in order to model the user.

The hat-mounted camera can be used to create a video database of the visual environment and visual experiences of the user. Standard techniques for video analysis might be used to classify different types of scenes and segment the video-stream into video-segments. However, the analysis of the video-stream from a hat-mounted camera can make use of the habits of humans. For example, humans tend to fixate onto objects and people of interest. Also the motion of objects relative to the user is a powerful indication of relevance of the object for the user. These and other facts makes the analysis of video from a first-person perspective very different to the standard problem of video-database analysis. By exploiting and being aware of humans habits, we are aiming to extract salient structure from the wearable's video-stream.

In order to create a system which continuously analyzes the visual environment of the user we have to consider at least the following points: First of all, the system has to be robust enough in order to handle data from a continuously moving camera under varying imaging conditions. Secondly, the calculations of the system should be simple enough in order to run in real-time or at least close to real-time. Thirdly, since no single feature will be capable to capture all objects which might be of interest, we have to allow the use of multiple features such as color, texture, motion and shape.

The following section describes a wearable system which is capable to recognize objects at a rate of 10Hz. The system can recognize in the order of 100 objects and has been shown to be robust to different lighting situations. Section 5 then proposes a system to automatically extract objects and object parts, calculate their motion and shape over time fully automatically using the hat-mounted wearable camera.

4 Recognizing objects with a wearable camera

An implementation of an augmented reality remembrance agent which does not use any sort of tags but a generic object recognizer in order to identify objects in the real world is a system called "Dynamic Personal Enhanced Reality System" (DyPERS). DyPERS retrieves 'media memories' based on associations with real objects the user encounters. These are evoked as audio and video clips relevant for the user and overlayed on top of real objects the user encounters. The system uses an audio-visual association system with a wireless connection to a desktop computer. The user's visual and auditory scene is stored in real-time by the system (upon request) and is then associated (by user input) with a snap shot of a visual object. The object acts as a key such that when the real-time vision system detects its presence in the scene again, DyPERS plays back the appropriate audio-visual sequence.

The audio-visual associative memory operates on a record-and-associate paradigm. Audio-visual clips are recorded by the push of a button and then associated to an object of interest. Subsequently, the audio-visual associative memory module receives object labels along with confidence levels from the object recognition system. If the confidence is high enough, it retrieves from memory the audio-visual information associated with the object the user is currently looking at and overlays this information on the user's field of view.

Whenever the user is not recording or associating, the system is continuously running in a background mode trying to find objects in the field of view which have been associated to an A/V sequence. DyPERS thus acts as a parallel perceptual remembrance agent that is constantly trying to recognize and explain – by remembering associations – what the user is paying attention to. Figure 2 depicts an example of the overlay process. Here, in the top of the figure, an 'expert' is demonstrating how to change the bag on a vacuum cleaner. The user records the process and then associates the explanation with the image of the vacuum's body. Thus, whenever the user looks at the vacuum (as in the bottom of the figure) he or she automatically sees an animation (overlaid on the left of his field of view) explaining how to change the dust bag. The recording, association and retrieval processes are all performed online in a seamless manner.

An important part of the system is the generic object recognizer, based on a probabilistic recognition system.

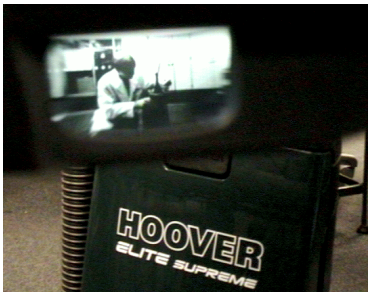

Figure 2: Sample Output Through heads-up-display (HUD)

Figure 3: A DyPERS user listening to a guide during a gallery tour

Objects are represented by multidimensional histograms of vector responses from local neighborhood operators. This object representation is considerably robust to view point changes. A probabilistic object recognition algorithm [21] is used, in order to determine the probability of each object in an image only based on a small image region. Experiments showed that only a small portion of the image (between 15% and 30%) is needed in order to recognize 100 objects correctly in the presence of viewpoint changes and scale changes. The recognition system runs at approximately 10Hz on a Silicon Graphics O2 machine using the OpenGL extension library for real-time image convolution. The vision system for DyPERS has been reported and evaluated in [22].

Obviously, the discrimination of 100 objects is not enough to be of practical use in an unconstrained real world scenario. However, by using information about the physical environment, including the location of the user, the time of day and other available information, the number of possible objects can be significantly reduced. Furthermore, information about the user's current interests further reduces the number of interesting objects.

5 Object learning model

The main limitation of the DyPERS system is that it needs direct input from the user in order to acquire object models. Rather than having a wearable system which needs supervision by the user we want to build a system which acquires object models in a un-supervised way. This section therefore proposes to build models of objects in the visual field of view by a three step procedure. Attentional objects can be extracted by tracking their position and motion in the visual field of view.

5.1 Three step learning of objects

This section describes the structure of the automatic object learning model. As pointed out earlier we are interested to extract objects and object parts from the visual environment. In order to extract a model of the objects we propose a three step learning procedure. This learning procedure is done in parallel for different features since no single feature is powerful enough to describe all possible objects in the environment.

The extracted object models contain information such as the color, the texture and the shape of the object parts. By analyzing and modeling the absolute motion and relative motion of different object parts we can group the different object parts to object hypotheses. If hypothesized objects are present in the visual field of the user repeatedly over time these hypotheses correspond very probably to real objects. The repetitive appearance of the objects also indicates the importance or relevance of the objects for the user in the visual environment.

Most computer vision algorithms rely on the fact that the camera conditions are not varying too much or that the varying conditions are known. In other words: Most computer vision algorithms are not robust enough to be used on a wearable computer. We should always keep this in mind, when we are using computer vision techniques in the context of wearable computers. However, we can make use of the fact that between consecutive frames, the imaging conditions do not change too much.

The object learning is done in three steps. First, for a short video sequence, consistent regions (e.g. in color and/or texture) are extracted and tracked over a number of frames. During tracking a representation of the image region is learned. This representation includes a model of multiple features: color, texture and shape of the region. This first learning step provides a list of region candidates which can be part of an object.

During the second learning step the algorithm identifies image regions which reoccur over a longer period of time. During this step the algorithm keeps track of relative arrangement (position and movement) of image regions in order to find sets of image regions which are consistently moving together. Such sets of image regions (and all possible subsets) are candidates to correspond to the same object. The output of the second step is therefore object hypotheses each consisting of a set of image regions and their respective spatial arrangement.

Keeping track of the relative position and movements of the different image regions is important to disambiguate which regions correspond to the same object.

The third learning step verifies the object hypotheses by continuously recognizing the hypothesized objects. If an object can be identified repeatedly in the visual field of view over an extended period of time that object hypothesis is stored permanently and considered as a learned object.

5.2 Attention Filtering

As pointed out earlier the system can exploit the first person view of the wearable system. Once objects have been extracted and hypothesized the system keeps track of the time the object is present in the visual field of view. During that time the system extracts a model of the object including color, texture and shape of the object. These models can then be used to recognize the repetitive appearance of the objects over extended periods of time. Objects which have been fixated several times by the user are very probably of high interest to the user.

Based on the above object models, the system can also extract the motion of the objects relative to each other and relative to the user. When the user visually tracks a moving object the object stays relatively still in the visual view the system whereas the background is constantly in motion. Such an event can be detected by extracting the motion of objects in the video stream.

Objects and people which reoccur in the field of view repeatedly over an extended period of time (e.g. over several days) the system can conclude that this object/person is of particular interest to the user. By extracting the hand of the user the system can also find out about physical interactions of the user with an object (e.g. touching or picking up) which indicate the relevance of such objects to the user.

All the above allows the system to do *attentional filtering* of the video-stream in order to extract objects which are of interest to the user of the wearable system. This particularity can be exploited in order to structure the video-stream in a way which is useful to the user.

5.3 Implementation Details and Examples

In our current implementation we are using k-means clustering [7] in order to extract candidate regions which might correspond to objects or object parts. More specifically, the system extracts color region by clustering pixels based on the standard color features U and V and the position X and Y of the pixel. The color features U and V are stable enough over small changes of lighting conditions in order to be able to match successive frames successfully. Using X and Y at the same time for clustering enables the extraction of compact clusters rather than clusters which might be distributed over the entire frame [9]. The first rows of figures 4 and 5 show the original color images (printed in black and white). The second rows of the same figures show the segmentation result of the color clustering.

As pointed out before, the system should use multiple features in order to be capable to model a wide variety of objects. Therefore, the system also extracts regions based on texture features. Currently we are using a second order autoregressive model [15] on a neighborhood of 5×5 pixels. We use three parameters of the autoregressive model[1] and the X and Y position of each pixel. The third row of figure 5 shows image regions which have been obtained by k-means clustering of the texture and position features.

As expected, image regions which have been extracted by clustering pixels based on color and texture, often correspond to objects or object parts. The first step of the object learning model consist of tracking such image regions over a couple of frames. In order to match clusters in successive frames we calculate the Euclidean distance between the mean value of the color features (or the texture parameter). For tracking the clusters over a short period of time (here over several seconds) we use a Viterbi algorithm.

The third row of figure 5 shows such a tracked sequence of color clusters (note that only every third frame of a 1 second, 10-frame sequence is shown). Interestingly, the extracted and tracked image region corresponds to a person standing in front of a wall. During using the wearable system in a test-run, this person happened to ask for directions and has been therefore focus of attention for a couple of minutes. During this time, a color and texture model of the person has been successfully extracted from the video provided by the hat-mounted camera. The fifth row of figure 5 shows the tracking of a cluster, which corresponds to a part of the wall behind the person in the video. Other parts of

[1] more specifically, we are using the mean and the first two coefficients. We are not using the forth parameter, namely the variance of the noise. See [15] for details

the same wall have been matched as well.

Figure 4 shows five frames of a six second video sequence. During that time, the user picked up a box, looked at it for about two seconds and put it back on the table. Therefore, the box has been an object of interest to the user. The third and the forth row show color clusters which have been successfully tracked over the entire sequence. Since these two clusters are moving together the system hypothesizes correctly that the two regions correspond to the same physical object. By tracking the hand of the user the system can extract the common motion of the object and the hand. This information enables the system to conclude that the user moved the object (here picking up). Information about objects of the visual environment and the interactions of the user with these objects might eventually be used to structure the video-stream according to relevance and importance to the user.

6 Conclusions

The paper proposes a wearable system using a hat-mounted camera in order to record and analyze the visual environment of the user. Several possibilities exist to model and recognize user context from video alone or with audio information combined. We have reported a system which used the hat-mounted camera successfully to recognize objects.

The paper also proposes the exploitation of human habits such as fixating objects of interest in order to structure and index the video data. The system extracts objects from the visual environment and segments interactions of the user with the environment. These can be used to structure the video data according to relevance and importance to the user. Video data structured in such a way can be of great use to remember and index events of importance.

7 References

[1] G. Abowd, A. Dey, R. Orr, and J. Brotherton. Context-awareness in wearable and ubiquitous computing. In *IEEE ISWC*, Oct 1997.

[2] H. Aoki, B. Schiele, and A. Pentland. Recognizing personal location from video. In *Workshop on Perceptual User Interfaces*, pages 79–82, Nov 1998.

[3] M. Bauer, T. Heiber, G. Kortuem, and Z. Segall. A collaborative wearable system with remote sensing. In *IEEE ISWC*, Oct 1998.

[4] S-F. Chang, Q. Huang, T. Huang, A. Puri, and B. Shahraray. *Advances in Mutimedia: Systems, Standards and Networks*, chapter Multimedia

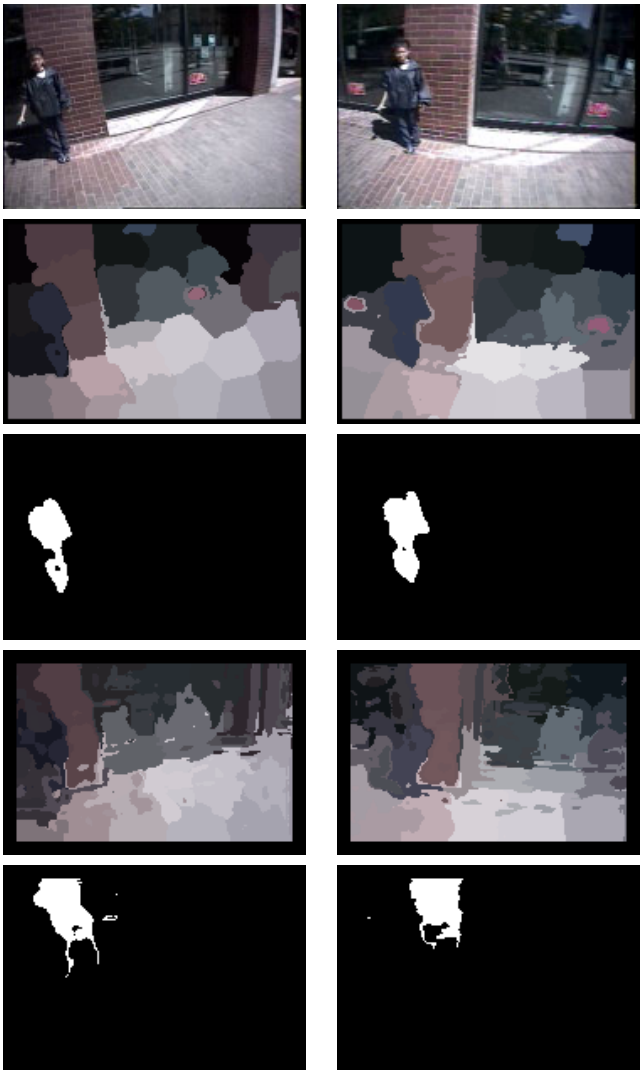

Figure 5: Object learning over 10 frames corresponding to 1 sec of video (first and last frames are shown). The first row shows the original color-image (in grey scale). The second row shows the color clusters and the third row shows a person which has been extracted based on color clustering and tracked over 10 frames. The forth row shows the texture clusters and the fifth row shows a part of the wall behind the person which has been extracted and tracked based on texture

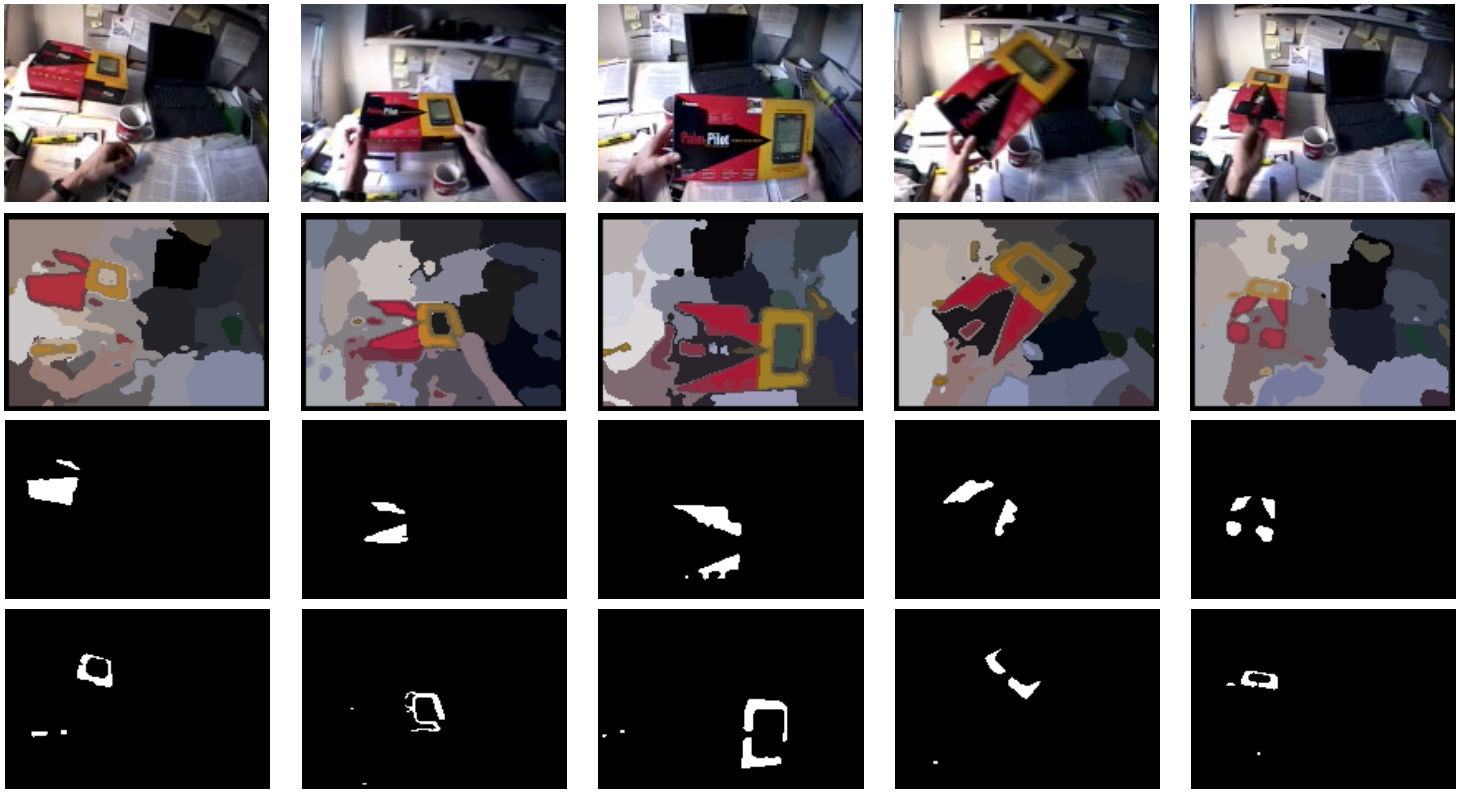

Figure 4: Object learning over 6 seconds of video (even though we are using 10 frames per second, only one frame per second is shown). The first row shows the original color-image (in grey-scale). The second row shows the color clusters extracted from the image. The third and the forth row show two parts of an object which has been picked up by the user and has been focus of attention to the user. Both parts of the object have been successfully tracked over the entire sequence

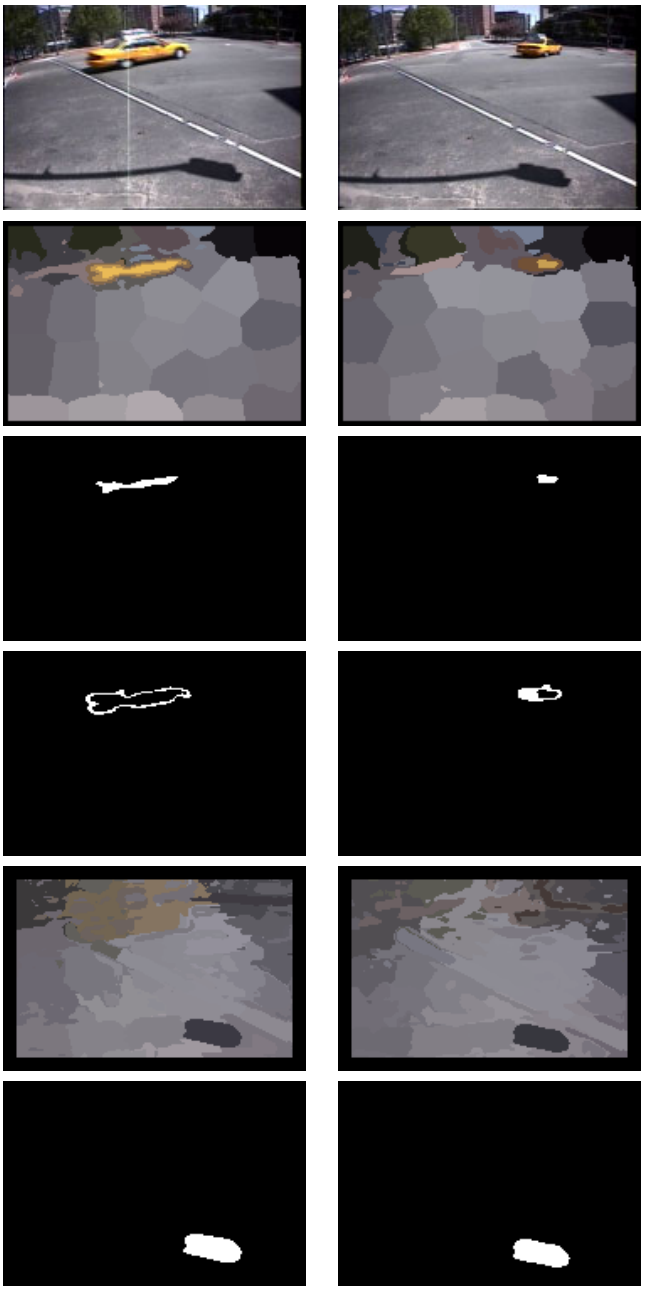

Figure 6: Object learning over 10 frames corresponding to 1 sec of video (first and last frame shown). The first row shows the original color-image (in grey-scale). The second rows shows the color clusters. The third and forth row show different parts of a taxi which have been extracted based on color clustering and tracked over 10 frames. The fifth row shows the texture clusters and the sixth row shows the a part of the street which has been extracted and tracked based on texture

Search and Retrieval. New York, Marcel Dekker, 1999.

[5] S.F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong. Videoq: An automated content based video search system using visual cues. In *ACM Mutimedia Conference*, Nov 1997.

[6] B. Clarkson and A. Pentland. Unsupervised clustering of ambulatory audio and video. In *Proceedings of the International Conference of Acoustics, Speech and Signal Processing*, 1999.

[7] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 1973.

[8] S. Feiner, B. MacIntyre, R. Höller, and A. Webster. A touring machine: Prototyping 3d mobile augmented reality systems for explorng urban environment. In *IEEE ISWC*, pages 74–81, Oct 1997.

[9] B. Heisele, U. Kreßel, and W. Ritter. Tracking nonrigid, moving objects based on color cluster flow. In *International Conference on Computer Vision and Pattern Recognition*, pages 257–260, 1997.

[10] R. Hull, P. Neaves, and J. Bedford-Roberts. Towards situated computing. In *IEEE ISWC*, Oct 1997.

[11] S. Kakez, C. Vania, and P. Bisson. Virtually documented environment. In *Proceedings of the First Intl. Symposium on Wearable Computers ISWC97*, Cambridge, MA, 1997.

[12] G. Kortum, Z.Segall, and M. Bauer. Context-aware, adaptive wearable computers as remote interfaces to 'intelligent' environments,. In *IEEE ISWC*, Oct 1998.

[13] M. Lamming and M. Flynn. Forget-me-not: intimate computing in support of human memory. In *Proceedings of FRIEND21 Intl. Symposium on Next Generation Human Interface*, 1993.

[14] S. Mann. Personal imaging and lookpainting as tools for personal documentary and investigating photojournalism. *ACM Mobile Networks and Applications*, 4:23–26, 1999.

[15] J. Mao and A.K. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25(3):173–188, 1992.

[16] J. Pascoe. Adding generic contextual capabilities to wearable computers. In *IEEE ISWC*, Oct 1998.

[17] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content based manipulation of image databases. *Intern. Journal of Computer Vision*, 18(3):233–254, 1996.

[18] R. Picard. A society of models for video and image libraries. *IBM Systems Journal*, 1996.

[19] J. Rekimoto, Y. Ayatsuka, and K. Hayashi. Augment-able reality: Situated communication through physical and digital spaces. In *IEEE ISWC*, Oct 1998.

[20] B. Rhodes and T. Starner. Remembrance agent: A continuously running automated information retreival system. In *Proc. of Pract. App. of Intelligent Agents and Multi-Agent Tech. (PAAM)*, London, April 1996.

[21] B. Schiele and J.L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR'96 Proceedings of the 13th International Conference on Pattern Recognition, Volume B*, pages 50–54, August 1996.

[22] B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system, dypers: Dynamic personal enhanced remembrance system. In *International Conference on Vision Systems*, Jan 1999.

[23] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *Second International Symposium on Wearable Computers*, Oct 1998.

[24] T. Starner, J. Weaver, and A. Pentland. A wearable computer based american sign language recognizer. In *First Intl. Symp. on Wearable Computing*, Cambridge, MA, 1997. IEEE Press.

[25] H-J. Zhang, S.W. Smoliar, J.H. Wu, C.Y. Low, and A. Kankanhalli. *Advances in Digital Libraries*, chapter Chapter 15: A Video Database System for Digital Libraries. Springer, Lecture Notes, 1995.