Adventurous Tourism for Couch Potatoes

Luc Van Gool[1,2], Tinne Tuytelaars[1], and Marc Pollefeys[1]

[1] University of Leuven, Kard. Mercierlaan 94, 3001 Leuven, Belgium
[2] Inst. Kommunikationstechnik, ETH, Gloriastr. 35, CH-8092 Zürich, Switzerland

Abstract. Two tourist guides are described. One supports a virtual tour through an archaeological site, the other a tour through a real exhibition. The first system is based on the 3D reconstruction of the ancient city of Sagalassos. A virtual guide, represented by an animated mask can be given commands using natural speech. Through its expressions the mask makes clear whether the questions have been understood, whether they make sense, etc. Its presence largely increases the intuitiveness of the interface. This system is described only very concisely.
A second system is a palmtop assistant that gives information about the paintings at the ongoing Van Dijck exhibition in the Antwerp museum of fine arts. The system consists of a handheld PC with camera and Ethernet radio link. Images are taken of paintings or details thereof. The images are analysed by a server, which sends back information about the particular painting or the details. It gives visitors more autonomy in deciding in which order to look at pieces and how much information is required about each. The system is based on image database retrieval, where interest points are characterised by geometric/photometric invariants of their neighbourhoods.

1 Introduction

The amalgam of more powerful computer vision techniques, faster computers, and wireless communications coupled to progressing miniaturisation yields technology that will soon pervade our lifes in many, so far mostly unimaginable ways. One area of applications probably is cultural tourism. One can imagine systems that allow us to visit sites only virtually and yet create a feeling of really being there. Alternatively, we may get on that plane and actually visit other places. But rather than spending our time on searching for the things we want to see and trying to get some information about them, virtual guides could accompany us and supply information 'just-in-time'.

In this paper, we describe two systems under development that can assist future tourists to get more information with less effort. Section 2 describes a system that supports virtual tours through the archaeological site of Sagalassos. Section 3 describes a palmtop guide to lead visitors through a real exhibition. The 1999 Van Dijck exposition in the Antwerp Museum of Fine Arts was taken as a show case. Section 4 concludes the paper.

F. Solina and A. Leonardis (Eds.): CAIP'99, LNCS 1689, pp. 98–107, 1999.

2 Virtual Sagalassos

The first system integrates a diversity of techniques to create virtual tours through the ancient city of Sagalassos. This old Greek city was destroyed by an earthquake in the 6th century. The goal of the demonstrator is to show this archeological site in 3D to a virtual visitor. He/She is accompanied by a virtual guide, who can be spoken to. The visitor can e.g. ask to go to a certain building or ask questions about finds that have been made at different places. The system integrates techniques for the 3D reconstruction of the site and its buildings from uncalibrated video sequences [6], techniques for the recognition of fluent speech (both Dutch and English, developed by our colleagues of speech recognition) and techniques to animate the face of the virtual guide. Missing parts of the buildlings are replaced by CAD-models composed by archeologists.

The guide makes the interaction more intuitive. A nod of the head shows that the system has "understood" the instruction, otherwise the eyebrows will be raised to prompt a repetition of the question. The visitor gets a 'no' if the instruction doesn't make sense at the current stage of the tour, e.g. if one asks to show a statue where there is none. Fig. 1 shows some example views during such a virtual visit. In the near future we plan to also let the guide's face talk, to initiate a real conversation with the visitor.

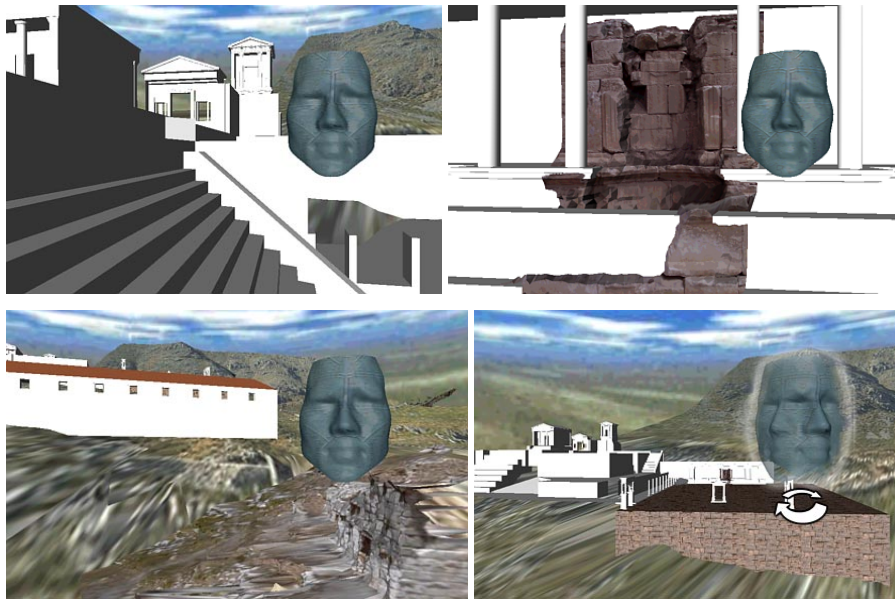

Fig. 1. *Holiday pictures from a trip to virtual Sagalassos.*

Such systems in a way yield an experience that is more complete then if one were on-site. It is e.g. possible to show the city in its diverse stages of development and to see the city grow towards its most prosperous age. Finds that are now stored in a museum can be put on their original locations. Ruins can be shown in their original context.

This sytem allows us remote access to a site. The system described next is meant as a guide to a real exhibition with the visitor on-site.

3 Palmtop Pictures of an Exhibition

3.1 System Overview

It is quite common for museums to provide tape or CDROM players as a replacement for human guides. Although they offer a kind of individual service in terms of the time a visitor can spend on different exhibits, it simultaneously is also more homogeneous, as no questions can be asked to get additional information. It would be better if a visitor could interact more dynamically with the virtual guide than pushing the 'next' or 'feedforward' button. An example would be pointing a camera at a piece to solicit information about it, and zooming in on parts if the visitor is intrigued by certain details. Such a system is the subject of this section.

The Museum of Fine Arts in Antwerp has for 1999 organised an important exhibition about the work of Antoon Van Dijck, in remembrance of the painter's 400th birthday. We have used this exhibition as a show case.

The basic setup is a handheld PC with built-in camera (e.g. Sharp's HC 4600 A) or a separate digital camera connected to the handheld pc via an infrared link. The digital image is then sent to a server via a wireless ethernet connection, where it is processed such that the correct information can be returned to the user. Figure 2 gives an overview. The server recognizes the painting the visitor

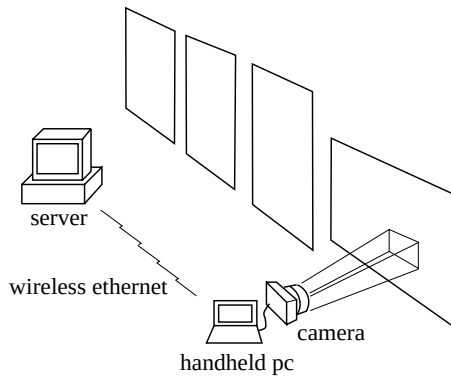

Fig. 2. *Overview of the system's architecture.*

is pointing the camera at. The visitor can also select details for more specific information. In fact, the server performs 'database retrieval'. Once the content of the picture has been recognized, the appropriate information is returned to the user over the same wireless ethernet connection.

As the visitor cannot be expected to always stand right in front of a painting and as illumination conditions may change (e.g. shadows of other visitors), the extracted image descriptors should remain invariant under such variability. Also, other visitors may occlude parts of the piece. This calls for the use of local features, i.e. features that are based on relatively small parts of the images. Section 3.2 describes the image processing steps in somewhat more detail.

An advantage of such a system - apart from its intuitive user interface - is that it can derive from the picture what the user is actually interested in and adapt the returned information accordingly. This is difficult to achieve with e.g. electronic tour guides that use the viewer's location as primary input (e.g. using infrared transmitters placed in the ceiling above each piece to be described) [1]. The visitor is also free to roam the museum and ask for information in whatever order, which is difficult with tape and CD-ROM players. Also, since all that is needed is a picture of the piece, no external adaptations have to be made. This makes the basic approach equally suited for less controlled environments like a moving exhibition, an open air show, or even a city tour.

Probably the most similar system is the Dynamic Personal Enhanced Reality System (DyPERS) of Jebara et al. [2], which has (among other applications) also been evaluated in a museum-gallery scenario. However, they recognize objects based on multidimensional receptive field histograms, which is probably less robust to changes in viewpoint than our system.

3.2 Recognizing a Painting or a Detail Thereof

The recognition process should be sufficiently robust, so that our camera–empowered visitor doesn't have to push aside other visitors to get shadow-free, frontal views. People may be walking in front of the camera causing partial occlusions, lighting conditions may change, etc. To deal with these problems retrieval is based on local, invariant features.

In a first step interest points are selected. For the moment these are only corners, detected with the Harris corner detector (i.e. points with a high degree of information content, as they are surrounded by intensity profiles that show large changes in different directions). Then, around each of these points a number of neighbourhoods are selected. The goal is to arrive at neighbourhoods that cover the same parts of the painting, irrespective of the viewpoint. This implies that the neighbourhood should deform as the observer moves around. Finally, from these neighbourhoods invariant features are extracted and used for interest point recognition in the database.

The system combines geometric and photometric invariance. Not only can the viewpoint change, but also the illumination conditions may vary or the spectral responses of the camera or scanner with which the database was produced may

differ from those of the visitor's camera. Both the construction of the neighbourhoods and the extraction of features from these neighourhoods must yield the same results irrespective of such changes. For this sytem, we consider invariance under 2D affine deformations

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

and linear changes in the 3 colour bands, where each band is subject to its own scaling and offset.

$$\begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} = \begin{pmatrix} s_R & 0 & 0 \\ 0 & s_G & 0 \\ 0 & 0 & s_B \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} o_R \\ o_G \\ o_B \end{pmatrix}$$

The crux of the matter are the invariant neighbourhoods, i.e. the definition of neighbourhoods that depend on the underlying image structure in such a way that their content is not altered by a change in viewpoint. The selected neighbourhoods correspond to viewpoint dependent parallelograms with corners as one of their vertices. When changing the position of the camera the neighbourhoods in the image change their shape so that they cover the same physical part of the object or scene. Figure 3 illustrates this. The image on the right is part of the database image "Ages of Man" (also shown in figure 5). The image on the left is part of a query image taken from a different viewpoint. Although the illumination conditions have changed drastically and the image is distorted due to the change in viewpoint, the affinely invariant neighbourhoods found in both images correspond, i.e. they are deformed in such a way that they still have the same physical content in spite of the change in viewpoint. Note that these neighbourhoods are constructed solely on the basis of each image separately.

This approach is reminiscent of the work of Schmid and Mohr [5] who also used interest points and their neighbourhoods, but in their case invariance is under Euclidean motions. Pritchett and Zisserman [4] also used parallelogram-shaped regions to find stereo correspondences under wide baseline conditions. In their case these were explicitly present in the image and not constructed from general surface textures as is the case here.

The steps in the construction of the affinely invariant neighbourhoods are all invariant under the forementioned geometric and photometric changes. More information on this construction is given elsewhere [7]. The geometric/photometric invariants used to characterize the neighbourhoods are moment invariants. For a complete classification of moment invariants of lower orders, we refer to [3]. The moments are so-called 'Generalized Colour Moments'. These better exploit the multispectral nature of the data. They contain powers of the image coordinates *and* the intensities of the different colour channels. They yield a broader set of features to build the moment invariants from and, as a result, moment invariants that are simpler and more robust to image noise.

Once such local, invariant descriptions have been derived, the image can easily be matched with the images in the database using a voting mechanism. For

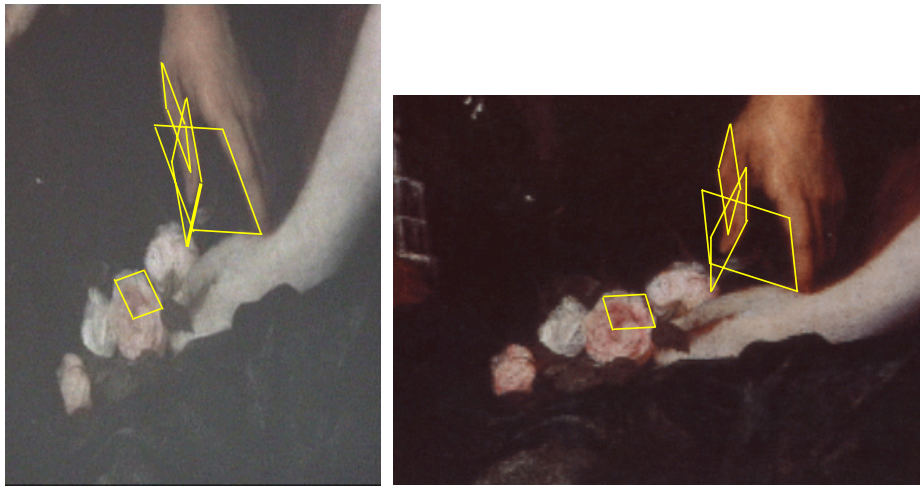

Fig. 3. *A detail of the database image "The Ages of Man"(right) and of the query image (left) with corresponding affinely invariant neighbourhoods.*

each neighbourhood, a feature vector is composed from moment invariants. The closest matching neighbourhood in the database is retrieved using the Mahalanobis distance. The painting with the maximum number of closest matches is selected. The efficiency of this voting mechanism is enhanced using hashing techniques. This renders the retrieval procedure less dependent on the image database size.

Each correspondence also gives an approximation for the affine transformation linking both images. This can be exploited in two ways. Firstly, this allows to reject false matches (i.e. matches that give a different affine transformation than the other matches). Secondly, it allows the server to precisely define what part of the database image the user is actually interested in, as the borders of the query image can be backprojected into the database image. The latter use is illustrated in the next section.

3.3 Experimental Results

Due to organisational problems as well as security considerations, we were in the end not able to evaluate our system at the real Van Dijck exhibition. Instead, we created a small mock-up gallery in our lab based on 50×70 cm^2 reproductions. It is important to note that this has complicated rather than simplified matters. The database images were scanned form a catalogue with high quality reproductions. Several posters in the mock-up gallery were of much lower quality, however, leading to a loss of colour. =2.6cm

Images of these posters were matched to high-resolution (1000×1500 pixels) digital images of 20 of Van Dijck's paintings stored in our database. An overview of the database is given in figure 4. In total, 12.978 affinely invariant

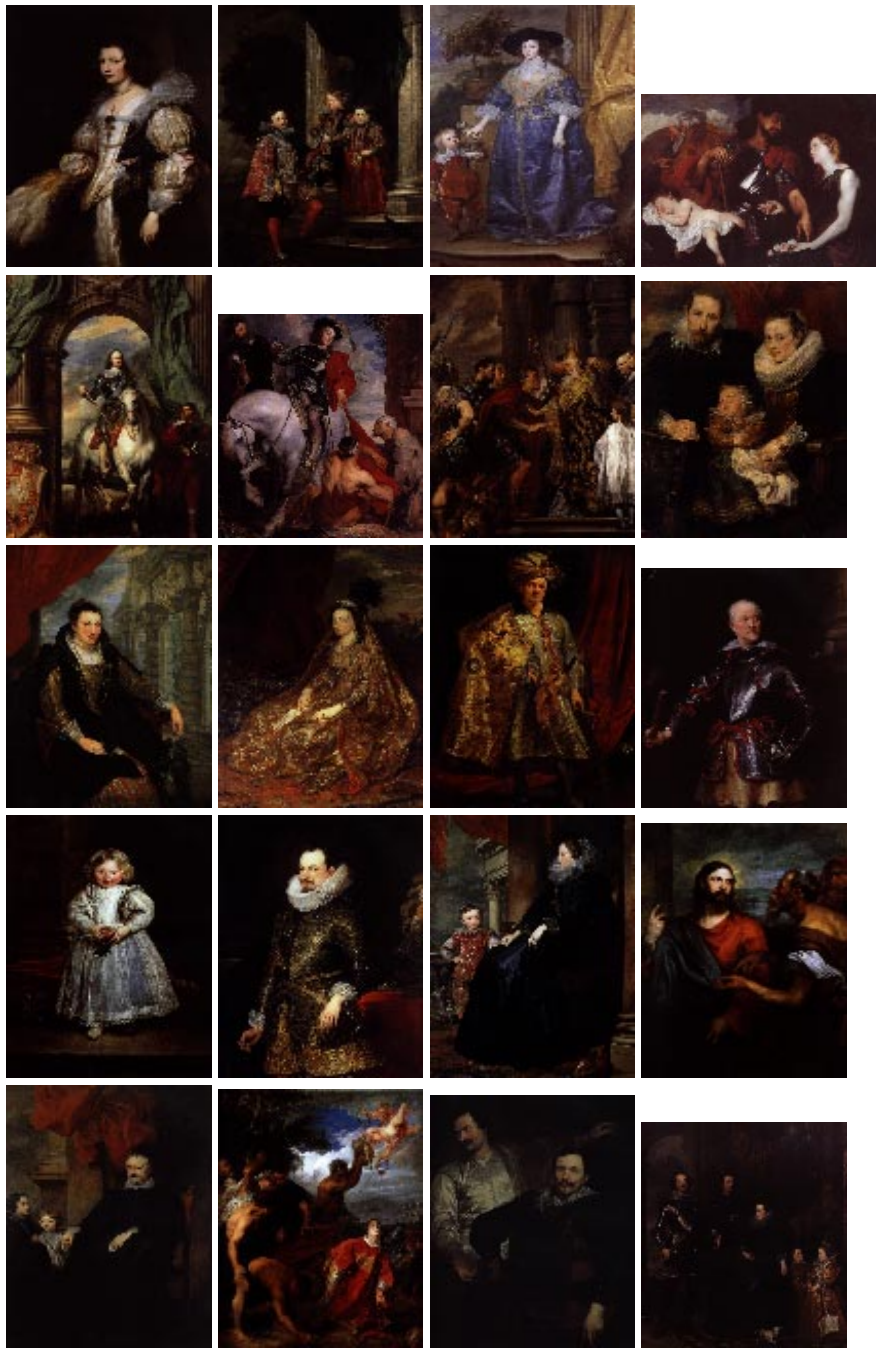

Fig. 4. *An overview of our database.*

neighbourhoods were found in these 20 images, i.e. an average of 650 regions per image.

As a first example, look at the painting called "The Ages of Man", shown in figure 5. At the top, the database image of the painting is shown with some possible user queries just below. Note the large differences in colour and the affine image deformations. Not visible in this rescaled figure is the difference in scale, which varied in between 80 and 130 percent. Nevertheless, the correct painting was retrieved from our database for each of the three query images. On top of the database image, parallellograms are added. These represent the corresponding image part as computed by the server, through the backprojection method mentioned earlier.

Similarly, a second example based on the painting "Maria Louisa de Tassis" is shown in figure 6.

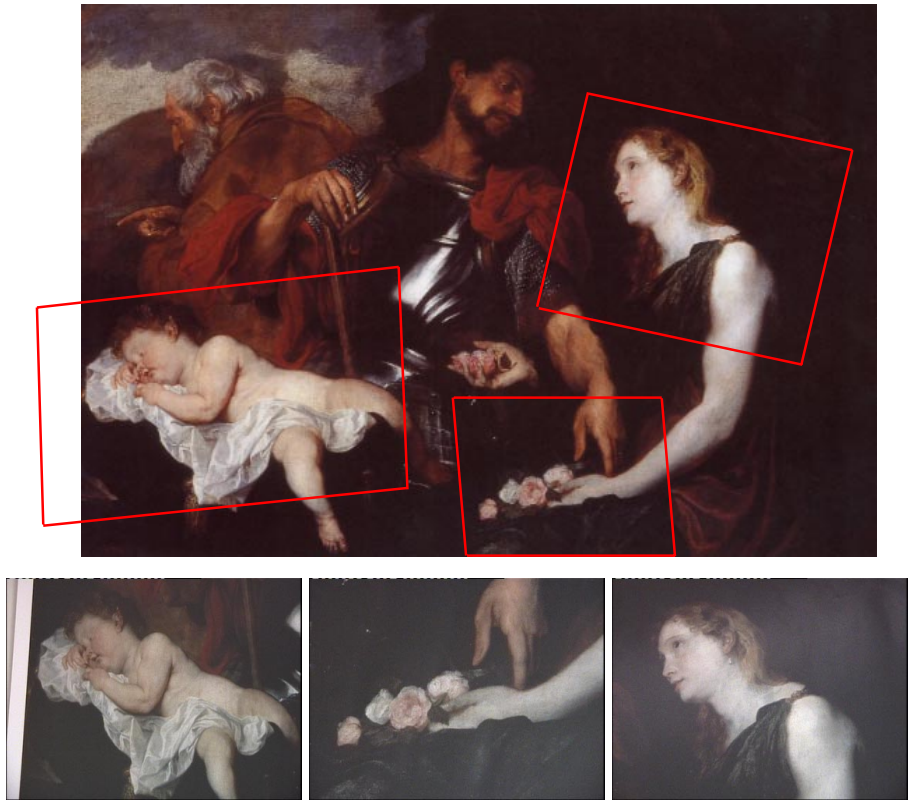

Fig. 5. *Given the query images shown below, the system was each time able to retrieve the correct database image (i.e. to identify the painting out of the list of 20) and to highlight the corresponding part therein (top).*

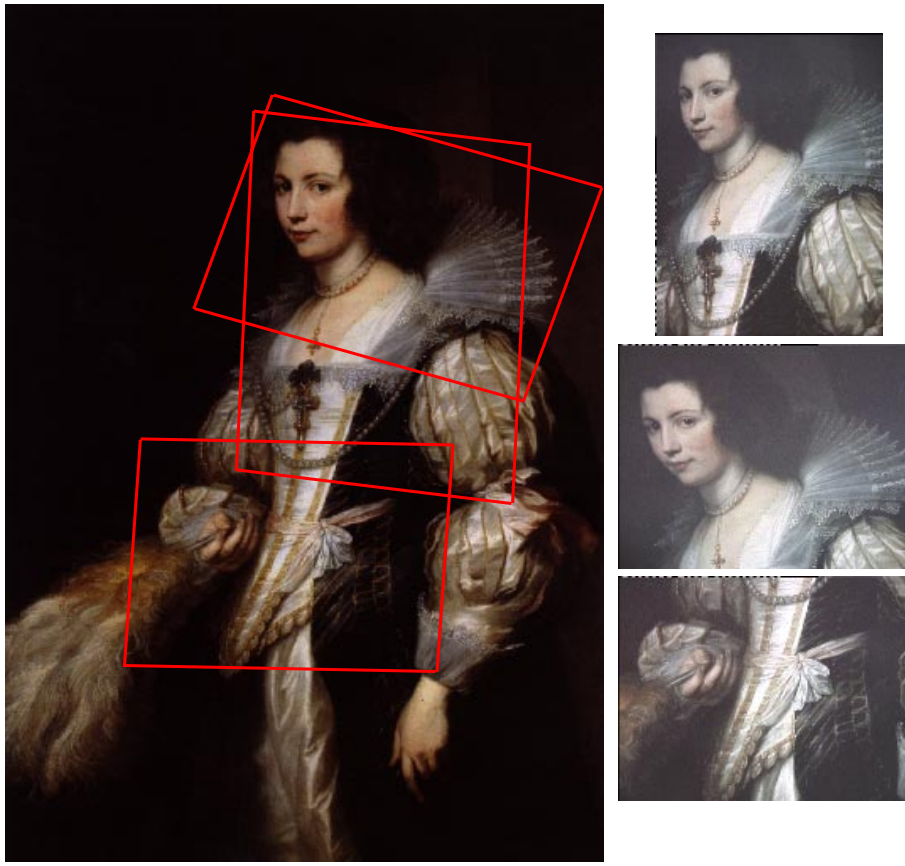

Fig. 6. *Given the query images shown at the right, the system was each time able to retrieve the correct database image and to highlight the corresponding part in the database image (left).*

4 Conclusion

Two tourist guides were described. The first one supports virtual tours through the archaeological site of Sagalassos based on a 3D reconstruction of the site. Mid-term goal is to fully integrate immersive visualisation with a natural dialogue with the animated virtual guide.

The second system describes a real tour through an exhibition. A visitor simply takes a picture of a piece s/he wants to learn more about. This picture is then transferred to a server via a wireless ethernet connection, where it is matched to the images in the database based on the notion of local, affinely invariant neighbourhoods. Based on the corresponding regions found, the correct information can be returned to the visitor. Advantages of the system are its easy and intuitive user-interface, its ability to respond to different levels of detail and

its flexibility (no external actions required). Although the experimental results presented in this paper are limited to images of 2D objects, the method may also be applied to 3D objects such as buildings or sculptures, as long as the surfaces can locally be approximated by planes [7].

Acknowledgements: Tinne Tuytelaars gratefully acknowledges an FWO grant of the Flemish Fund for Scientific Research. Implementations were done in TargetJr. The authors gratefully acknowledge the support by the IUAP 4/24 project IMechS, financed by the Belgian OSTC.

References

1. B. Bederson *Audio Augmented Reality: A prototype Automated Tour Guide*, ACM Human Computer in Computing Systems conference (CHI'95), pp. 210-211, 1995.
2. T. Jebara, B. Schiele, N. Oliver, A. Pentland *DyPERS: Dynamic Personal Enhanced Reality System* MIT Media Laboratory, Perceptual Computing Technical Report nb 463.
3. F. Mindru, T. Moons, L. Van Gool *Color-based Moment Invariants for the Viewpoint and Illumination*
Independent Recognition of Planar Color Patterns, to appear at ICAPR, Plymouth, 1998.
4. P. Pritchett, A. Zisserman *Wide baseline stereo matching*, Proc. International Conference on Computer Vision (ICCV '98), pp. 754-759, 1998.
5. C. Schmid, R. Mohr *Local Greyvalue Invariants for Image Retrieval*, PAMI Vol. 19, no. 5, pp 872-877, may 1997.
6. M. Pollefeys, R. Koch and L. Van Gool, *Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters*, Int. Conf. on Computer Vision, pp. 90-95, 1998.
7. T. Tuytelaars, L. Van Gool *Content-based Image Retrieval based on Local Affinely Invariant Regions*, International Conference on Visual Information Systems, Visual99, 1999.