Proceedings of Mirage 2005, INRIA Rocquencourt, France, March, 1-2 2005

A MOSAIC-BASED VISUAL MEMORY WITH APPLICATIONS TO ACTIVE SCENE EXPLORATION

Birgit Möller and Stefan Posch

{moeller,posch}@informatik.uni-halle.de
AG Pattern Recognition & Bioinformatics, Institute of Computer Science
Martin-Luther-University Halle-Wittenberg, 06099 Halle / Saale, Germany

ABSTRACT

Processing visual data is an important ability of interactive systems to act in dynamically changing environments. Looking at the human visual and cognitive system this requires efficient mechanisms for data processing and storage as well as intelligent strategies for data acquisition. In this paper we present a visual memory supporting efficient representation of image sequences of active cameras in an online fashion. The memory is based on mosaic images extending the field of view of a camera in space and time. As one prototypical field of application for the memory active scene exploration is discussed. The temporally and spatially integrated data of the memory combined with additional feature maps serves as an ideal starting point for the selection of focus points. Hence the memory demonstrates the combination of efficient acquisition and storage of visual data and helps to provide interactive systems with high flexibility to operate in dynamically changing environments.

1. INTRODUCTION

Acting in dynamically changing environments is a great challenge for interactive and mobile systems. They have to be designed flexible enough to cope with unknown situations. Scene understanding and behaviour planning is usually achieved by analyzing data acquired with available sensor devices and interpreting these data based on world knowledge. Especially visual devices yield an important source of information for scene analysis and understanding, providing data in terms of single images and complete sequences. The latter ones contain static as well as dynamic information, however, at the same time suffer from large, redundant data volumes that hamper efficient analysis.

Human beings strongly rely on visual data processing when analyzing their environment and planning new behaviours. The human visual system includes a variety of structures to efficiently memorize data of different semantic levels and is based on selective, attention-driven data acquisition. This allows for efficient information analysis despite limited processing capabilities. The goal to achieve high flexibility in interactive systems based on visual data processing thus implies to supply systems not only with efficient storage and analysis mechanisms for sequences but also to combine these with active view point selection to guide data-driven scene exploration.

In this paper we present our concept of a visual scene memory that meets both requirements. It is based on *mosaic images* supporting storage-efficient representations of image sequences. All images are warped towards a common coordinate system and are then integrated into a single frame by fusing their color values thus removing redundant information. We aim at representing data of stationary rotating and zooming cameras. Since our intented field of application for the memory is given by interactive systems mosaicing should be done in an *online fashion*. The data represented in the mosaic needs to be accessible at any point in time during image acquisition to not obstruct the interactivity of the systems. Besides, interactive or mobile systems usually provide only limited ressources for computation and storage rendering impossible the processing of complete sequences simultaneously. Hence we immediately integrate each new image into the evolving mosaic overcoming the restriction to buffer the images.

The coordinate frame of our memory is defined based on a *polytope* regularly arranged around the optical center of the camera. It is represented with a set of tiles each equipped with a local euclidean coordinate system. This enables the direct application of conventional image analysis techniques to the iconic data and thus simplifies data access, e.g., compared to cylinders or spheres often used to represent mosaic images [1, 2]. In contrast to these approaches usually working offline (e.g., [3, 4]), iconic data of the monitored scene is available immediately after acquisition and integration of each new frame and without need for explicitly rendering parts of the polytope.

Easy online update of the memory data is achieved facilitating an additional image plane, *the focus image plane*. It augments the polytopial memory structure and allows for efficient data update despite discontinuities between different tiles. Finally, adequate representation of image data including different levels of zoom is enabled by nesting differently scaled instances of a polytope. According to the tiled memory structure and the resolution hierarchy we call these enhanced mosaic images *multi-mosaics*.

Dynamic scene data yields important clues for scene interpretation and understanding. However, independently moving objects hamper image registration and cause blurring in mosaic image integration as they do not conform to the global motion model imposed. Detecting and masking these objects in mosaicing avoids this but results in complete loss of dynamic information. The visual memory described here represents dynamic as well as static information. While static scene parts are integrated in the multi-mosaics data structure, dynamic parts are simultaneously extracted and represented separately in a graph data structure. It may serve as a base in subsequent processing steps but also allows deriving clues for interesting events in a scene that are worth to be explored in detail.

One possible area of application for the memory is active scene exploration. Many approaches have been published (e.g., [5, 6]) to simulate human visual attention in artificial systems. While most work relies on the analysis of single image data (in some cases enhanced with motion information from difference images), our visual memory allows to extend view point selection to spatially integrated data. Selecting new focus points, image data may be considered that is no longer in the view of the camera but nonetheless of importance. Equipped with additional feature maps the memory supports the representation of interest measures of certain regions of a scene, thus it is well-suited to combine data storage and its active acquisition.

The remainder of this paper is organized as follows. Section 2 introduces our multi-mosaic data structure. In section 3 algorithms to extract and represent dynamic data are discussed, while section 4 outlines the support of the visual memory for active scene exploration. The paper finishes with results in section 5 and a conclusion.

2. MULTI-MOSAIC IMAGES

A mosaic image is built from all images of a sequence warping them into a common coordinate frame. Defining this frame, the degrees of freedom of the camera, the area of application for the mosaic images and the intended way of processing the data have to be considered. We aim at representing the complete field of vision of stationary rotating and zooming cameras in an online fashion. Each new image should be registered and integrated immediately into the evolving mosaic image without temporally storing the data. Besides, euclidian coordinates should be provided to support direct analysis of the mosaic data applying conventional image processing techniques.

One widely used coordinate frame for representing mosaic image data is defined by single planes. However, projecting image data acquired with stationary rotating and zooming cameras onto a single plane results in large distortions rendering the mosaic image unemployable for further analysis. Cylinders or spheres yield a straightforward approach to represent image data from stationary rotating and zooming cameras avoiding distortions [2]. However, explicitly representing spheres is more difficult, particularly if not all images ever registered should be kept (cf. [1, 4]). Online registration of new images also yields a great challenge using spherical representations since image registration is most comfortable adopting image planes. Finally, using a spherical or cylindrical coordinate frame violates for example the fundamental invariance of colinearity, which is explicitly or implicitly assumed by a wide range of image processing techniques. Hence performing image analysis, e.g., edge detection or even more elaborate tasks like object recognition based on regions, would require to develope new algorithms suitable to cope with non euclidean coordinates.

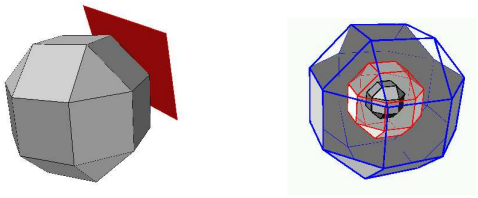

Figure 1: *Polytopial coordinate frame (rhombicuboctahedron) with focus image plane attached (left) and sketch of the hierarchical structure of the visual scene memory (right).*

2.1. Polytopial Coordinate Frames

To overcome the problems outlined above our visual memory relies on a coordinate frame defined by a set of various tiles. These are arranged regularly around the optical center of the camera. Their exact arrangement is derived from *polytopes* approximating a sphere (Fig. 1, left) and hence allowing to minimize distortions while at the same time providing euclidean coordinates for image processing. Each tile is located tangentially to the sphere and equipped with a local euclidean coordinate system. The origin of the 3D coordinate frame of the polytope is located at the optical center of the camera. Its z-axis is aligned with the optical axis of the camera on acquisition of the first image of the sequence for convenience. The scaling of the polytopial tiles is defined as the focal length of the camera. It is extracted facilitating an offline calibration strategy providing a functional mapping between hardware parameters and corresponding focal length which is used in online generation. Self-calibration techniques that might be used alternatively for online estimation of the focal length have proven to be too unstable in long-term online mosaicing.

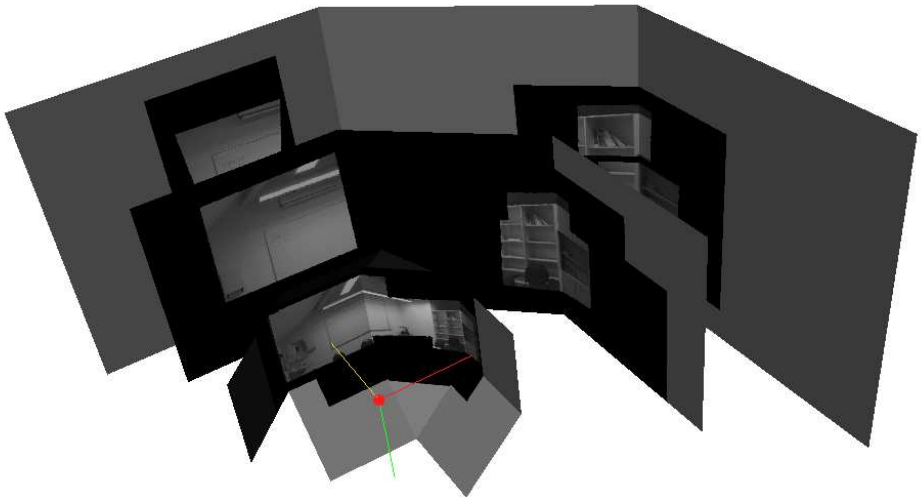

Figure 2: *Multi-mosaic representation of image data acquired with varying camera zoom settings.*

2.2. Multi-Resolution Representation

Image sequences often contain image data acquired with different camera zoom settings. Usually, such image sequences are integrated into a single mosaic image covering only a single level of resolution. However, this is not adequate since data acquired with higher resolution has to be downsampled and low-resolution data has to be interpolated. This shows, that image data with varying granularity demands to use a hierarchy of differently scaled mosaic images for suitable representation. We meet this requirement by nesting tiles derived from differently scaled instances of a polytope into each other (Fig. 1, right). They cover a discrete set of resolutions. As new image data becomes available the one instance is chosen for data representation that best meets the focal length of the input image (Fig. 2). Since most of the time not all parts of a scene have to be represented simultaneously and in all resolutions, the polytope instances are only partially represented yielding a *sparse memory representation* [7].

2.3. Online Generation and Update

Mosaic images are generated estimating parameters $\vec{p}$ of a motion model $T_{\vec{p}}$ for each image of a sequence *(registration)*. The motion model describes the camera motion and hence allows for its compensation. All images are warped towards the common coordinate frame and then *integrated* fusing their color information. Both steps can be performed offline or online. While in offline processing all images are registered and integrated simultaneously, we generate our multi-mosaics in an *online* fashion. Each image is registered and integrated as it is acquired into the evolving mosaic image yielding a continuous data update. However, the resulting representation is optimal only locally since registration and integration are solely based on the current input image and the mosaic image itself which results in accumulation of errors over time. Nevertheless such a strategy is favorable especially if the mosaic image generation should be performed on systems providing only limited storage and processing capabilities as it is usually the case with mobile interactive systems.

2.3.1. Registration

The motion of stationary rotating and zooming cameras can exactly be modelled with homographies. Their parameters are estimated using *projective flow* [8]. The algorithm is based on local optical flow computations, globally constraint by the projective motion model. It is implemented in an iterative framework utilizing piecewise linear homography approximations. A resolution hierarchy (which must not be mixed up with the hierarchical memory representation outlined in section 2.2) is applied to account for large displacements between images. Since parameter estimation in levels of low resolution still requires a minimum of image quality the resolution hierarchy cannot be applied to arbitrary large displacements. Hence, camera data for rotation and focal length is additionally used to derive an initial estimate for the parameters. This leads to improved robustness and faster convergence and at the same time allows for larger rotations between successive frames. Even non-overlapping frames can be registered to the mosaic, provided suitable reference data in the multi-mosaic representation. Error accumulation is reduced by estimating parameters in *frame-to-mosaic-mode* [9]. Each image is registered using a suitable clip of the mosaic. Hence all images formerly integrated at least implicitly take part in the current estimation step yielding improved image quality without processing all images of a sequence simultaneously.

2.3.2. Integration

Image integration is accomplished by selecting a single input image as source for each mosaic pixel. Selection is based on the time-stamps of the input images. Each pixel is assigned the value from that input image providing the most recent data. This strategy leads to a segmentation of the mosaic image into regions each originating from a different image of the sequence. Visible seams might result between neighboring regions in case of changing lighting conditions or camera exposure settings. Hence linear or sigmoid blending functions are applied in a small neighborhood along region boundaries to smooth transitions.

2.3.3. Focus Image Plane

Both in registration as well as in integration the structure of the polytopial coordinate frame is well-suited with exception of discontinuities between different tiles. Explicitly accounting for them in registration and integration is time-consuming, hence we adopt an additional image plane to simplify processing even more. A *focus image plane (FIP)* is attached to the polytope (Fig. 1) tracing the movements of the camera and its focal length. It augments the underlying polytopial structure defined by the arrangement of the single tiles and is used as reference in registration and integration. Image data is only projected onto the polytope itself if the position and orientation of the FIP require an update. This happens if the camera orientation has changed significantly from that one represented by the FIP and hence new image data is projected outside its domain. Since the FIP is chosen to be two to three times larger in size than a single input image usually several images can be registered and integrated before updates become necessary. Image data is then projected onto the related tiles. Subsequently position and orientation of the FIP are updated and new reference image data is back-projected from the polytope to the new FIP.

3. DYNAMIC SCENES

Mosaicing algorithms compensate for the camera motion during acquisition of an image sequence. As outlined parameters of a suitable motion model are estimated from differences between subsequent images based on the assumption that these differences are exclusively caused by camera motion. This assumption does not hold in dynamic scenes where differences between images are induced by the camera motion as well as by independently moving objects. Their motion is not covered by the global motion model and can hamper registration and integration. We cope for dynamic parts of a scene by detecting independently moving objects and masking them from registration and integration. This yields a mosaic representation including only static scene parts. However, dynamic data is often most relevant for scene analysis and interpretation. Hence we additionally track moving objects, extract their dynamics and represent the data in a supplementary graph data structure. It serves a suitable base for data analysis but also allows verification of motion data and consistency of static data represented in the multi-mosaic memory.

3.1. Motion Detection

Various approaches are known in the literature for detecting differences between registered images. They range from relatively simple intensity residual calculations and algorithms based on normal flow [10] up to statistical [11] and elaborate optical flow based methods [12]. Within the context of this work we primarily intend to detect independently moving objects without exactly describing their motion by means of appropriate motion models. Hence we adopt residual-based approaches. In previous work [13] we also employed normal flow and convergence analysis [14], however, these techniques did not result in improved performance. Hence detection of independently moving objects within a sequence is accomplished by calculating normalized pixelwise intensity residuals $R_I(x,y)$ between registered images I_{ref} and I_ω. We define the reference image I_{ref} to be the clip of the mosaic image already used for registration. I_ω denotes the current input image I warped to the coordinate frame of I_{ref} applying the estimated projective transformation $T_{\vec{p}}$. Residuals are averaged in a neighborhood N of each pixel (x,y) to account for image noise:

$$\begin{aligned} R_I(x,y) &= \frac{1}{C} \sum_{\substack{(x',y')\in \\ N(x,y)}} \left(\frac{I_{ref}(x',y')}{\bar{I}_{ref}} - \frac{I_\omega(x',y')}{\bar{I}_\omega} \right)^2 \\ C &= |N(x,y)| \ , \ I_\omega = T_{\vec{p}}\,(I) \end{aligned}$$

The residuals are normalized with respect to the average intensities $\bar{I}_{ref}$ and $\bar{I}_\omega$ in the images to reduce the influence of varying image energies. The resulting residuals are thresholded using an empirically chosen threshold θ_R yielding a pixelwise motion map C_{θ_R} (Fig. 3(c)):

$$C_{\theta_R}(x,y) = \begin{cases} 0, & \text{if } R_I(x,y) \le \theta_R, \quad \text{(static)} \\ 1, & \text{otherwise} \qquad\quad \text{(moving)} \end{cases} \tag{1}$$

Large residuals indicate image locations that were not registered by the global motion model and most of the time result from independently moving objects.
Pixels detected as moving are masked during integration and for registration steps of subsequent frames yielding a robust sequence mosaicing even in case of large moving objects [15]. However, since moving scene parts are initially unknown a proper initialization is required. We

(a) reference mosaic

(b) new frame, already registered to the mosaic coordinate system

(c) pixelwise motion map

(d) resulting mosaic image with moving pixels omitted in integration

Figure 3: *Residual-based detection of an independently moving booklet at the bottom of the new frame. Shifting an object from one position to another yields two moving regions. Neither its former position nor the current one match the data represented in the mosaic image.*

assume that there are either no or only small moving objects present for the first two images of the sequence to be processed. Otherwise an initial motion map has to be provided externally.
Figure 3 shows two registered images and the resulting map of pixels classified as moving. The object shifted from one position to another yields two compact groups of moving pixels: Neither its current position nor the former one match the image data represented in the mosaic and thus introduce large residuals. As moving pixels are masked during integration the moving object is not integrated in the mosaic image as intended in our approach. However, its old position is kept in the mosaic representation of the static scene background although the object has already disappeared and became part of the dynamic scene foreground. This obviously would lead to an inconsistent data representation. Hence such inconsistencies are eliminated performing an additional analysis of object dynamics as described in the remainder of this section.

3.2. Component Tracking

Masking independently moving objects during integration yields mosaic images restricted to static scene parts. At the same time all dynamic information of objects gets lost. To overcome this limitation we extract dynamic data by tracking moving objects over time (more details can be found in [13]). This is accomplished by first applying morphological operations to the motion maps $C_{\theta_R}(x, y)$ and then segmenting regions from this data. Neighboring regions according to small point distances are grouped into connected components to account for variance in segmentation and tracked over time. Small regions with an area below a certain threshold are omitted. Matching between successive frames is performed based on intensity histograms, size and position data. All components of both images are matched pairwise against each other. If connected components are left without a valid match they are decomposed into sub-components and matched again. In doing so even disintegration and merging of moving objects can be detected.
Since no model assumptions about moving objects are included, only rigid or slow deforming objects can be tracked by this approach. We prefer this approach to not limit tracking of objects to special classes applying explicit models. We note that in any case *all* objects can be detected and masked for robust mosaicing as discussed in the preceeding subsection.

3.3. Analysis of Trajectories

Tracking connected components yields trajectories specifying their positions over time defined in terms of the centroid positions. These trajectories are represented in a data structure called *correspondence graph*. Nodes are associated with matched connected components and subcomponents. Edges indicate matches and are labeled with pairwise match distances. Each node contains as additional information component specific data like the duration of tracking and the variance of centroid positions within the last frames. The latter one enables to check the consistency of the motion data. Spurious object positions, e.g., the old position of the formerly static booklet in Fig. 3 (cf. section 3.1), are tracked over time but do not show significant changes of position. Thus, the variance of centroid positions is very small and allows for identification of these situations. The corresponding components can be marked as non-moving and the mosaic representation is updated by integrating image data which was initially masked from integration. In this way data consistency of the mosaic image is guaranteed with a small shift in time according to the number of frames needed for variance analysis. This is also true for cases where objects change between static and moving, as described: Analysis of tra-

jectories allows toggling between an interpretation of the obejct being part of the static background or belonging to the dynamic scene foreground.

4. ACTIVE SCENE EXPLORATION

Active cameras provide interactive systems with a large flexibility in visual data acquisition. However, large data volumes of the image sequences impede easy extraction. Exploiting redundancies, the amount of data can be reduced representing the sequences using mosaic images as described. Additionally, already during acquisition the data volume can be kept small by selective acquisition. Most of the time not all parts of a scene are equally interesting for exploration. Rather only a few regions are worth to be scanned in detail. Hence guiding attention to those interest points helps to reduce the overall amount of data the system has to cope with.

One question raising in selective data acquisition is the *"Where to look next"*-problem. Many proposals published in this area answer this question by simulating mechanisms of human visual attention. Human attention is triggered by a variety of different features like color, symmetry, size, texture, contrast or motion [16]. All features are individually weighted according to underlying goals. In artificial systems such mechanisms are for example simulated based on features maps and/or neural networks (e.g., [17, 18]). In most approaches the image data used to select a new focus point is restricted to the current input image. Sometimes this data is combined with motion data from residual images and maps masking formerly focused points in a scene. However, human attention is not restricted to analyzing the current sensor input. Rather memorized data is equally important. Human beings often guide attention to scene points that are currently not in their field of view but were some moments before. Such points often provide important information to solve a given task. This implies to make use of spatial integration in artificial systems and to avoid restricting focus point selection to data of the current input image. Integrating visual data allows to search for focus points directly on the memorized data without need for expensive hardware-based re-explorations of a scene. Zooming to a given focus point reduces the current field of view of a camera. Keeping formerly acquired data in memory supports detailed data acquisition without loss of global context and relations. A mosaic-based memory representation as presented in this paper yields a well-suited starting point for efficient view point selection in active scene exploration.

To focus on the general idea how to employ the multi-mosaics for scene exploration we use a simple measure based on entropy combined with motion data. More sophisticated approaches can, however, easily be integrated in a straightforward way. In the next subsection we outline the incorporation of additional features maps into the multi-mosaic structure, while the remaining subsections discuss interest measures and coarse as well as detailed exploration strategies.

4.1. Feature Maps

The multi-mosaic data structure outlined in section 2 is expanded with additional polytopes to support active scene exploration. The first one represents time-stamps indicating the last point in time the region has been in the field of view of the camera (called *visit*). The other polytopes store different local interest measures of the mosaic. Points in the scene are regarded worth to be focused if they receive a large interest index compared to other points in the scene and have not been in the camera's field of view for some period of time.

The polytopes are scaled to a coarse resolution thus data is represented for groups of neighboring pixels (usually sized 15×15) instead of single pixels. Their resolution is fixed which is primarily due to memory efficiency and has proven sufficient for our experiments. Online data updates are performed analogously and in parallel to updates of the image data on the polytope using additional focus planes.

4.2. Coarse Exploration Triggered by Entropy

As mentioned we use local entropy as a single interest measure. This yields regions with large contrast and a wide variety of different pixel values as interesting. The local entropy $E(x, y)$ is calculated in a neighborhood N of a pixel (x, y):

$$E(x,y) = \sum_{v=0\ldots255} \frac{1}{p_v} \log p_v \quad (2)$$

$$p_v = \frac{1}{|N(x,y)|} \sum_{(x,y)\in N(x,y)} \delta_{I(x,y),v} \quad (3)$$

A new focus point f is selected based on the interest measure $E(x, y)$ weighted with the time elapsed since last visit t_l:

$$f = \underset{(x,y)}{\operatorname{argmax}}\, M(x,y) = \underset{(x,y)}{\operatorname{argmax}}\, E(x,y) \cdot \frac{(t - t_l)}{\gamma} \quad (4)$$

γ denotes a scaling factor controlling the period of time after which an explored focus point may capture full attention again. Pixels in a rectangular neighborhood around the current focus point are explicitly masked in subsequent exploration steps and their interest measures are only gradually increased again (black rectangles in figure 4). This prevents the algorithm from continuously fixating one single region with high entropy. Once a new

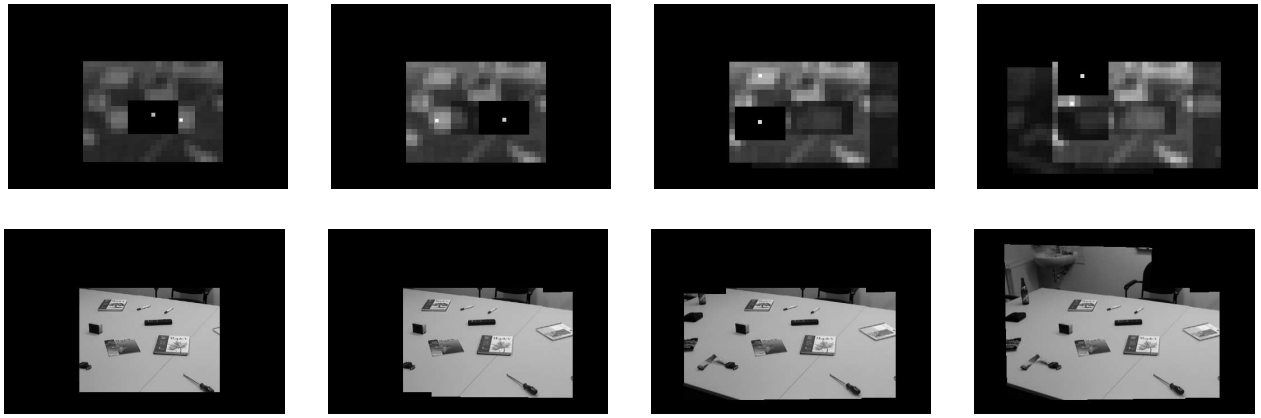

Figure 4: *Top: Attention maps evolving over time; Bottom: corresponding mosaic images. Bright intensity values indicate regions of high interest. Selected focus points are marked by white squares while black rectangles result from masking the currently focused region in subsequent exploration steps.*

focus point has been selected the camera moves towards it. Since large rotations between successive frames induce increased risk of registration failure, they are divided into smaller moves registered robustly one after the other.

4.3. Detailed Exploration Triggered by Entropy

Scanning a scene in one single resolution is not sufficient for most analysis tasks. While large parts of a scene are adequately represented in coarse resolution each scene usually includes regions of special interest that are worth to be represented with more details. Such regions are for example defined by specific objects. Our exploration algorithm accounts for such regions by exploring them in detail. Interesting regions are selected based on thresholding local entropy. If a focus point is selected for fixation and its entropy exceeds the empirically chosen threshold the camera first focuses the point and then initiates zooming. Currently a fixed number of zoom steps is performed before exploration in coarse resolution continues. Automatically determining the number of necessary steps to adequately focus the region of interest can be achieved by analyzing local structure or texture information.

4.4. Detailed Exploration Triggered by Motion

Dynamic data yields important clues for scene understanding. Especially addition and removal of objects are of great interest since these events indicate changes in the scene structure and provide valueable clues for intepretation of the visual data. Moving objects cause differences between the current image and the reference mosaic data and thus are detected by our motion detection algorithms (sec. 3). In particular, objects inserted as well as objects removed from the scene yield moving regions that are actually static and do not show variance of their centroid positions as described in section 3.3. Detection of such regions is included in the computation of interesting points and as a consequence these regions immediately capture attention. As for detailed exploration triggered by entropy the system now switches to zooming mode to get details about the changes. Subsequently active exploration continues by searching for a new focus point in coarse resolution using entropy. To account for robust motion analysis exploration is limited to small saccades in case of moving pixels detected within the field of view of the camera.

5. RESULTS

The visual memory including its prototypical application to active scene exploration have been tested on various image sequences of a meeting room (see images on the left of figure 5 for an example). The scene is dominated by a table with a couple of different objects. Figure 4 illustrates the first four steps of a coarse exploration triggered by entropy. The top row contains attention maps evolving over time. In the bottom row the corresponding mosaic images are depicted. Bright intensity values indicate large interest measures $M(x, y)$. New focus points selected according to these measures are marked by white squares in each of the maps. Black rectangles indicate the neighborhood of the currently selected focus points masked for further exploration. The algorithm starts by subsequently guiding the camera to the three books on the left. Due to their highly textured covers entropy values are larger than for other objects in the scene. However, as time passes by other objects become equally interesting due to temporal weighting. Hence the algorithm is guided to explore regions of high entropy first, but to fixate regions with low interest measures later on as well.
In figure 5 a snap-shot of another exploration session is

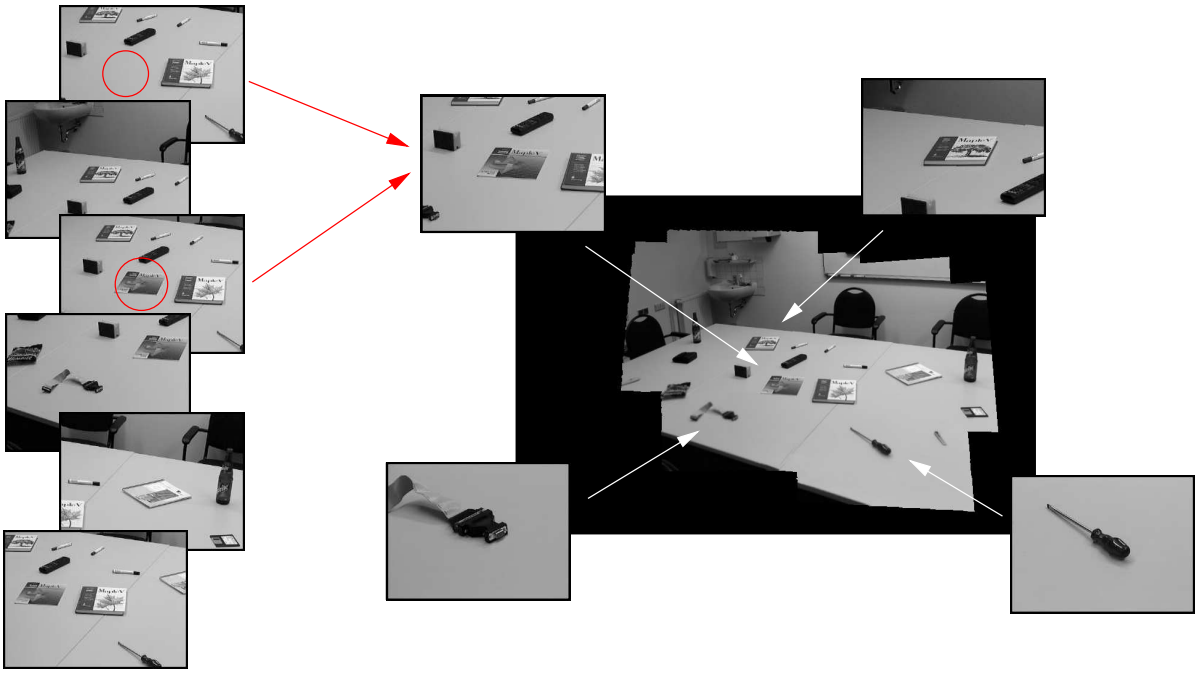

Figure 5: *Mosaic image with some regions automatically explored in detail by zooming in. Regions were selected either based on high local entropy or motion (top left region). On the left, some images of the complete sequence are shown.*

shown, where some interest points have been explored in detail by zooming in. Three of these regions were selected due to local entropy exceeding a certain threshold. The forth region (magnified in the top left corner of the mosaic image) was selected due to motion. As obvious from the image sequence depicted at the left hand side of the figure, the book was initially not present in the scene and added during exploration. It was detected as moving by our motion detection algorithms and tracked over time. However, after the book has been positioned on the table the variance of its centroid position vanished. The (formerly) moving region was classified as "spurious" indicating changes in the scene structure. Hence the corresponding region was regarded highly interesting for detailed exploration. Furtheron corresponding image data was integrated into the mosaic representation to recover consistency with the static scene background again.

6. SUMMARY AND CONCLUSION

Designing interactive systems operating in dynamically changing environments is a challenging task. Especially the analysis of visual data is an important ingredient towards this goal. In this paper a visual memory was presented accounting for efficient representation of image sequences acquired with active cameras. The memory is based on mosaic images extending the field of vision of cameras and providing spatially and temporally integrated data representations. These so called multi-mosaics are generated in an online fashion providing access to the represented information immediately after the integration of new data. The topological structure of the memory is defined by polytopes yielding adequate coordinate frames for image data from stationary rotating and zooming cameras. Euclidian coordinate frames on single tiles support online registration and integration, as well as the application of standard image analysis algorithms directly to the data. A resolution hierarchy of polytope instances allows adequate representation of visual data acquired with different camera zoom settings.

As one prototypical field of application for the visual memory active scene exploration is discussed in this paper. The memory's extended field of vision yields a well-suited base to efficiently support selection of focus points. In contrast to other approaches based on analyzing exclusively the current sensor input for focus point selection our memory supports to search the complete volume of visual data ever acquired and memorized. To this end the visual memory is equipped with additional polytope instances storing time-stamps and interest measures. Not all parts of a scene are equally interesting. Rather only some few points are worth to be scanned in detail. Our algorithm accounts for this by automatically zooming in to regions with exceptional high entropy or regions where changes in the scene structure were indicated due to analysis of motion trajectories.

To conclude, the visual memory structure presented in this paper demostrates how to efficiently memorize and process image sequences of active cameras. The extended field of vision of mosaic images yields a well-suited base for efficient active scene exploration and thus yields clues how to efficiently combine data storage and acquisition in building flexible interactive systems.

7. REFERENCES

[1] G. Bishop and L. McMillan, "Plenoptic modeling: An image-based rendering system," in *SIGGRAPH Computer Graphics Proceedings*, August 1995, pp. 39–46, Annual Conference Series.

[2] S. Coorg and S. Teller, "Spherical mosaics with quaternions and dense correlation," *Int. Journal of Comp. Vision*, vol. 37, no. 3, pp. 259–273, 2000.

[3] H. S. Sawhney, S. Hsu, and R. Kumar, "Robust video mosaicing through topology inference and local to global alignment," in *Proc. of European Conf. on Computer Vision*, 1998, pp. 103–119.

[4] H.-Y. Shum and R. Szeliski, "Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment," *IJCV*, vol. 36, no. 2, pp. 101–130, February 2000.

[5] S. Egner and C. Scheier, "Feature binding through temporally correlated neural activity in a robot model of visual perception," in *Proc. of Int. Conf. on Artificial Neural Networks*, 1997, pp. 703–708.

[6] Y. Yeshurun, "Attentional mechanisms in computer vision," in *Artificial Vision, Human and Machine Perception*, V. Cantoni, Ed. 1997, pp. 43–52, Plenum Press, New York.

[7] B. Möller, D. Williams, and S. Posch, "Towards a mosaic-based visual representation of large scenes," in *Proc. of 6th Open German-Russian Workshop*, Russian Federation, Aug. 2003, pp. 108–111.

[8] S. Mann and R.W. Picard, "Video orbits of the projective group: A new perspective on image mosaicing," Tech. Rep. 338, MIT Media Laboratory Perceptual Computing Section, 1996.

[9] P.J. Burt and P. Anandan, "Image stabilization by registration to a reference mosaic," in *Image Understanding Workshop*, 1994, pp. (1):425–434.

[10] M. Irani, B. Rousso, and S. Peleg, "Computing occluding and transparent motions," *Int. Journal of Comp. Vision*, vol. 12, no. 1, pp. 5–16, 1994.

[11] H.S. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 18, no. 8, pp. 814–830, August 1996.

[12] A. Bruhn, J. Weickert, and C. Schnörr, "Combining the advantages of local and global optic flow methods," in *Pattern Recognition (24th DAGM Symposium)*, Luc von Gool, Ed. 2002, vol. 2449 of *LNCS*, pp. 454–462, Springer.

[13] B. Möller and S. Posch, "Detection and tracking of moving objects for mosaic image generation," in *Pattern Recognition, Proc. of 23rd DAGM Symposium*. 2001, LNCS 2191, pp. 208–215, Springer.

[14] M. Ben-Ezra, S. Peleg, and B. Rousso, "Motion segmentation using convergence properties," *ARPA Image Understanding Workshop*, pp. 1233–1235, November 1994.

[15] B. Möller, D. Williams, and S. Posch, "Robust image sequence mosaicing," in *Pattern Recognition, Proc. of 25th DAGM Symposium*, Magdeburg, 2003, LNCS 2781, pp. 386–393, Springer.

[16] J.M. Wolfe, "Visual attention," in *Seeing*, De Valois KK, Ed., pp. 335–386. Academic Press, San Diego, CA, 2. edition, 2000.

[17] M. Bollmann, R. Hoischen, and B. Mertsching, "Integration of static and dynamic scene features guiding visaul attention," in *Mustererkennung*. 1997, pp. 483–490, Springer.

[18] J. Steil, G. Heidemann, J. Jockusch, R. Rae, N. Jungclaus, and H. Ritter, "Guiding attention for grasping tasks by gestural instruction: The gravis-robot architecture," in *Proc. of IEEE Conf. on Intelligent Robots and Systems (IROS)*, 2001, pp. 1570–1577.