Visual Scene Memory Based on Multi-Mosaics

Birgit Möller and Stefan Posch

Institute of Computer Science, Martin-Luther-University Halle-Wittenberg,
06099 Halle/Saale, Germany,
{moeller,posch}@informatik.uni-halle.de,
WWW home page: http://www.informatik.uni-halle.de/~posch/AG/

Abstract. Visual data acquired with active cameras yields an important source of information for interactive systems. However, since image sequences usually comprise large data volumes and notable portions of redundant information analysis is often diffcult. Hence, data structures are required that allow for compact representation of image sequences. In this paper we introduce our concept of a visual scene memory. The memory is based on mosaic images enabling compact image sequence representation by fusing all sequence images into one single frame while eliminating redundancies. Since interactive systems put special demands on mosaicing techniques we developed a new mosaic concept called *multi-mosaics* well-suited to be used with interactive systems. The memory is focussed on adaquate representation of iconic data, however, not restricted to it. Rather higher-level data, particularly motion data as well as data suitable for active camera control are additionally included completing the visual scene representation.

1 Introduction

Visual data is one of the most important sources of information for interactive and mobile artificial systems. Active acquisition of this data enables these systems at least in principle to autonomously act in dynamically changing environments and to perform intuitive interactions with human communication partners. However, interactive and especially mobile systems usually accommodate only for limited resources to store and process data. As a consequence, it is not possible for these systems to store and process all redundant image data acquired by an active camera. Rather sophisticated mechanisms for efficient data selection and storage are required that enable the systems to gather visual data and delay analysis as needed by later requirements.

Image sequences contain different kinds of information, dynamic as well as static data. Additionally the data implicitly cover different levels of abstraction, ranging from pure iconic information to intermediate-level primitives like edges or corners, and finally to semantic information like object recognition results. As the level of abstraction increases, also the compactness of structures used to represent this knowledge increases. However, the more abstract the data is the more limited is its applicability (e.g. 3D data for robot navigation or specific features for object recognition purposes).

In this paper we present our concept of a visual scene memory for representing *iconic* multi-resolution image data. The basis for this memory is given by *mosaic images* that enable efficient representation of image sequences acquired with active cameras. Such a memory supports a wide variety of possible areas of application due to its unspecialized, low-level data representation. However, since mosaic images extend a camera's field of view in space as well as in time it is straightforward to enhance the pure iconic representation with additional higher-level data that also might benefit from an extended field of view. Thus, our visual memory also supports representation of higher-level data like motion data as well as feature maps for autonomous scene exploration suitable to control active cameras. This yields a scene representation covering different levels of abstraction.

Using a visual scene memory based on mosaic images with interactive systems puts special demands on the algorithms used. On the one hand it is essential to support online data integration and easy data updates. Further on, an easy to use interface for data access is required which in particular has to support the application of conventional image analysis techniques directly to the data. Fulfilling these requirements has led to the development of a new mosaic image concept called *multi-mosaics* that enable interactive systems to efficiently represent and analyse image sequence data acquired with active cameras.

2 Mosaic Image Basics

Mosaic images are a widely used approach for efficient representation of image sequence data acquired with active cameras. The basic idea is to warp all images of a given sequence into a common coordinate frame applying suitable transformations *(registration)*. Subsequently one single mosaic image is constructed from all warped images fusing their color information *(integration)*. A mosaic image thus extends a camera's field of view in space and time and allows to eliminate redundancies within a given sequence. Consequently, the data volume of a sequence is significantly reduced when represented in terms of a mosaic image and hence data storage as well as analysis is notedly simplified.

Image sequence registration is usually based on a suitable mathematical model for the camera motion. It allows to describe changes between subsequent images of a sequence induced by the camera motion. The complexity of possible models mainly depends on the degrees of freedom of the camera and on scene structure. In our framework we use stationary but rotating and zooming cameras. Movements of such cameras can be described by a projective motion model. Although such cameras enforce mobile systems to stay at a fixed position within a scene during data acquisition, most of the time scenes can adequately be modeled by acquiring image data from a few "key positions" within a scene. Thus, it is usually not necessary to allow arbitrary camera movements which usually cannot be modeled by closed form transformations at all.

The motion of stationary rotating and zooming cameras can be described using homographies with 8 dofs. During registration, for each image of a sequence parameters for this model are estimated that allow to warp the image

into the common coordinate frame. In our system parameter estimation is accomplished with the *perspective flow* approach [1]. It is based on optical flow computations restricted by the projective motion model. For image integration new image data is essentially copied region wise to the final mosaic image. To smooth discontinuities along region boundaries appropriate blending functions are applied.

3 Multi-Mosaics

As already outlined, using mosaic images with interactive and particularly mobile systems enforces special constraints on the mosaicing algorithms that exclude many existing approaches to be applied directly to this new area of application. Primarily, mosaicing has to be done in *online mode*. Due to limited resources of mobile systems the complete image sequences cannot be stored and processed, as e.g. proposed in [2] or [3]. Rather it is necessary to register and integrate each new image immediately as it becomes available to overcome the need for storing all sequence images explicitly.

A second important aspect when representing mosaic images is to choose an appropriate reference frame the sequence images are warped into. Common choices for such frames are for example a single plane, a cylinder or a sphere. The later two choices allow for adequate and distortion free representations of image data acquired with rotating cameras as used in our approach. With regard to interactive systems, however, representing image data of rotating cameras in spherical coordinates has drawbacks like singularities when representing the complete viewing sphere and absence of collinearity. Since the vast majority of existing image analysis algorithms depends on Euclidean coordinates they cannot be applied to mosaic data projected onto spheres. This would severely restrict possible areas of application for the memory. Thus, our approach is based on *polytopes* that yield piecewise planar approximations of a sphere and, hence, reduce distortions while at the same time providing Euclidean coordinates. Additionally a mosaic consists of a set of differently scaled polytopes nested into each other to account for adequate representation of multi-resolution data resulting from a zooming camera. According to the current focal length of the camera the polytope instance is chosen for data projection that minimizes scaling effects. The resulting visual memory data structure consisting of multiple planes and multiple levels of resolution is called a *multi-mosaic image* (Fig. 1).

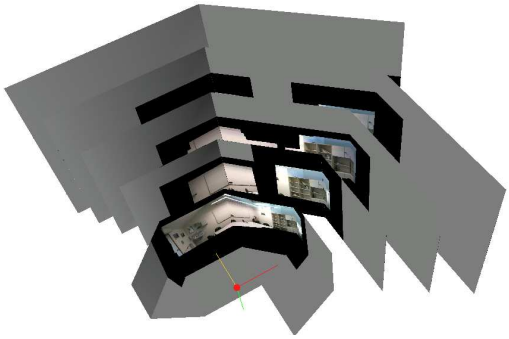

Fig. 1. Exemplary multi-mosaic: image data is projected onto a polytopial coordinate frame minimizing distortions while providing Euclidean coordinates.

Besides providing Euclidean coordinates multi-mosaics also support efficient online mosaicing. Although the piecewise planar tiles already enable easy registration and integration of new data we adopt an additional plane, the so called *focus image plane*, to further improve the handling of the memory structure and to minor the influence of discontinuities between neighboring tiles. The focus plane is attached tangentially to the polytope (Fig. 2). New image data is directly registered and integrated into this plane, hence polytope access is omitted. The focus plane traces the camera trajectory and its position and orientation is updated if the camera orientation differs too much from its current orientation. Only in these situations image data is copied into the multi-mosaic data structure. Hence, the focus plane serves as some kind of mediator between input data and memory. It stores the most recent data for direct access while the polytope itself yields a longer-term iconic memory.

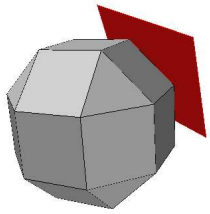

Fig. 2. Polytope with focus image plane attached.

4 Extensions to Higher-Level Data

The multi-mosaics provide an efficient iconic representation of image sequences. They yield a large flexibility in data analysis by supporting the direct application of existing image analysis algorithms. Nevertheless, the representation can be further improved by additionally providing data structures to include higher-level data resulting from intermediate processing steps as well. Employing the extended view of the multi-mosaics to represent these data allows for more flexibility analyzing image sequences and finally leads to better exploition of available data to improve the capabilities of interactive systems. Our implementation is currently focussed on representing motion information as well as data for guiding active scene exploration as outlined below. However, other kinds of data can easily be included as well, and preliminary work in this direction has already been done to include object recognition results.

4.1 Motion Data

One kind of higher-level data important for scene analysis and understanding is motion data. Detection of independently moving objects not covered by the global motion model yields the base for extracting dynamic data contained in image sequences and, thus, is of high importance for scene understanding. In addition, detecting these movements is important for registration and integration since they often detoriate parameter estimation and cause integration errors.

To handle moving objects motion detection and tracking algorithms are thus included in our memory. Independently moving objects are first detected computing intensity residuals. Subsequently moving pixels are masked from integration and subsequent registration steps. In addition they are segmented into regions and connected components which are tracked over time to extract the trajectories of moving objects (Fig. 3).

Temporal correspondences of connected components and related trajectories are then represented in an additional data structure, the *correspondence graph*. Besides encoding the trajectories of moving objects it also allows to derive rudimentary interpretations of scene data [4].

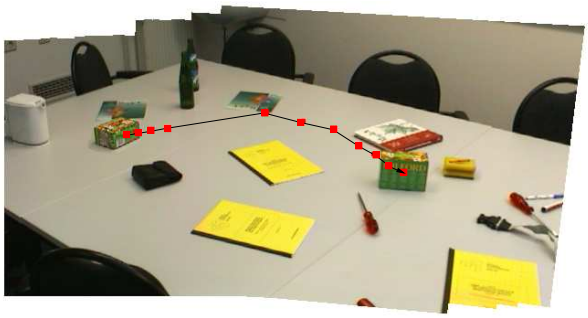

Fig. 3. Representation of higher-level data: moving objects are represented in terms of their trajectories included within a correspondence graph data structure.

4.2 Active Camera Control: Scene Exploration

One important question that has to be answered adopting active cameras for data acquisition is how to control the camera's movements. Often camera control is done simulating mechanisms of human visual attention and scene exploration [5]. In doing so focus points are automatically selected according to local interest measures calculated on the current input image. Compared to these single-view approaches mosaic images yield a richer source of information for focus point selection. Their temporally and spatially extended field of view facilitates to exploit all visual data of the scene acquired so far. Therefore, the attention mechanism allows the system to also explore scene parts not currently visible which is not feasable without an iconic memory. Using an iconic memory may also be of importance if the relevance of a given measure may vary in the course of time due to changing requirements, for example via user interaction, and computing all conceivable measures of interest in advance is prohibitive due to their potential number. Given the multi-mosaic the measures may be computed on demand when the information is actually required. Our multi-mosaic memory also supports the representation of interest measures. This is accomplished with an additional polytope for each interest measure to be represented. Currently local entropy and motion information are used for focus point selection. However, due to the Euclidean coordinate system of the multi-mosaics arbitrary image processing algorithms can be applied to the data allowing for flexible feature extraction.

Figure 4 shows an example mosaic image automatically acquired by active scene exploration. The clips enlarged were automatically chosen to be explored in detail by appropriate camera zoom. The selection was based on high entropy and motion information. This example demonstrates that the new multi-mosaic concept is well-suited not only to support efficient iconic and even higher-level representations of image sequences acquired with active cameras but also to support autonomous active scene exploration. Thus, the visual memory supports interactive systems with an integrated framework for image sequence representation and active data acquisition.

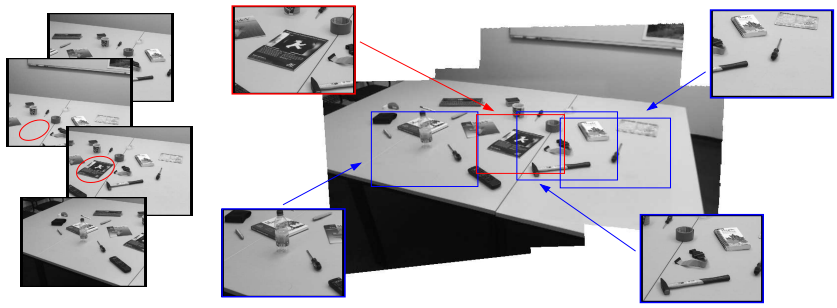

Fig. 4. Mosaic image captured by autonomous scene exploration. The image clips show regions automatically selected for detail exploration by camera zooming due to high entropy (blue boxes) or motion (red box) as can be seen from the images on the left.

5 Conclusion

The visual memory presented yields a well-suited approach to efficiently represent image data acquired with active cameras. The iconic representation supports a wide variety of possible areas of application, in particular mobile robots that perform interactions with humans will benefit from such a visual memory [6]. In this setup, the mobile system acquires multi-mosaic images from different positions while waiting for request from the human communication partner. The mosaics yield a rich source of information that can be exploited to solve specific tasks according to future demands, e.g. object learning. Thereby the memory is not restricted to pure iconic data but also supports the representation of higher-level data. Such a strategy facilitating a visual memory is superior to collecting data always "just-in-time" when it is actually required, and thus significantly improves the flexibility of interactive systems acting in everyday life environments.

References

1. Mann, S., Picard, R.: Video orbits of the projective group: A new perspective on image mosaicing. Technical Report 338, MIT Media Laboratory Perceptual Computing Section, Boston, USA (1996)
2. Bishop, G., McMillan, L.: Plenoptic modeling: An image-based rendering system. In: Proc. Int. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH), Los Angeles, CA (1995) 39–46
3. Shum, H.Y., Szeliski, R.: Systems and experiment paper: Construction of panoramic image mosaics with global and local alignment. Int. Journal of Computer Vision **36** (2000) 101–130
4. Möller, B., Posch, S.: Analysis of object interactions in dynamic scenes. In: Pattern Recognition, Proc. of DAGM Symp. LNCS 2449, Schweiz, Springer (2002) 361–369
5. Wolfe, J.: Visual attention. In De Valois, K., ed.: Seeing. 2. edn. Academic Press, San Diego, CA (2000) 335–386
6. Möller, B., Posch, S., Haasch, A., Fritsch, J., Sagerer, G.: Interactive object learning for robot companions using mosaic images. In: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, Edmonton, Alberta, Canada (2005) to appear.