DYNAMIC ATTENTIVE SYSTEM FOR OMNIDIRECTIONAL VIDEO

Iva Bogdanova, Alexandre Bur, Pierre-André Farine

Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Microengineering (IMT)
Pattern Recognition Laboratory (PARLAB)
Rue A.-L- Breguet 2, 2000-Neuchâtel, Switzerland

ABSTRACT

In this paper, we propose a dynamic attentive system for detecting the most salient regions of interest in omnidirectional video. The spot selection is based on computer modeling of dynamic visual attention. In order to operate on video sequences, the process encompasses the multiscale contrast detection of static and motion information, as well as fusion of the information in a scalar map called saliency map. The processing is performed in spherical geometry. While the static contribution collected in the static saliency map relies on our previous work, we propose a novel motion model based on block matching algorithm computed on the sphere. A spherical motion field pyramid is first estimated from two consecutive omnidirectional images by varying the block size. This latter constitutes the input of the motion model. Then, the motion saliency map is obtained by applying a multiscale motion contrast detection method in order to highlight the most salient motion regions. Finally, both static and motion saliency maps are integrated into a spherical dynamic saliency map. To illustrate the concept, the proposed attentive system is applied to real omnidirectional video sequences.

***Index Terms*—** dynamic visual attention, block matching on the sphere, omnidirectional image processing

1. INTRODUCTION

To perceive the environment efficiently, the human visual system (HVS) proceeds by selecting salient targets. The targets are explored successively by means of saccadic eye movements, which are responsible for shifting the fovea onto the current fixated target. Defined as selective attention, this mechanism can be seen as a preprocessing step, which reduces the amount of information that will be processed later by the brain.

The purpose of any dynamic visual attention model is to provide an automatic selection of potential regions of interest all over the sequence duration. The selection process relies on motion as well as static feature contrasts. It encompasses the feature extraction from the video sequence and its integration to define the resulting dynamic saliency map. This scalar map indicates salient locations, in the form of a saliency distribution. At the end, most salient regions of interest are defined from the saliency map using a selection process based on a neural network.

Several investigations focusses on the architecture of computer models of dynamic visual attention [1, 2, 3]. In order to operate on video sequences, such models generally integrate an additional motion component to the classical saliency-based model that has been proposed in [4]. All these models operate in the Euclidean geometry, i.e. they fit images obtained with conventional cameras. Applying directly such models on omnidirectional images leads to inaccurate deformations [5], due to the non-planar geometry of the image. Specific mappings, like panoramic or log-polar mappings, attempt to reduce somehow the distortions but they do not succeed completely.

Nowadays, the development of applications involving omnidirectional imaging is increasing because of its larger field of view and it is widely used in robotics and surveillance. In addition, visual attention is an attractive solution to reduce complexity issues in computer vision applications. Indeed, it can be conceived as a preprocessing step which allows a rapid selection of a subset of the available sensory information. Once selected, the salient targets become the specific scene locations on which higher level computer vision tasks can focus. Therefore, the development of an attentive system operating on omnidirectional video may lead to prospective computer vision applications.

In this paper, we propose a dynamic attentive system for detecting the most salient regions of interest in omnidirectional video. All the processing is performed in the spherical geometry in order to avoid the omnidirectional image distortions. In fact, it was shown in [6] that there exists an equivalence between the central catadioptric projection and the two-step mapping onto the sphere. Therefore, the approach is applicable to any omnidirectional image that can be mapped on the sphere.

The paper is organized as follows. Section 2 presents the dynamic attentive system that operates in spherical geometry, including the static model (Subsection 2.1), the motion model

The authors would like to acknowledge the support of the Swiss National Center for Competence in Research IM2 and the Swiss National Science Foundation under Contract FN-108060.

(Subsection 2.2) and the integration of both models (Subsection 2.3). Then, Section 3 illustrates the concept by showing a few results on (spherical) omnidirectional video sequences. Finally, we conclude and give some basic future work outlines in Section 4.

2. DYNAMIC VISUAL ATTENTION MODEL ON THE SPHERE

A computational visual attention model dedicated to video sequences must consider both static and motion features. The model is therefore composed of two parts: (i) static saliency map issued from a set of static features and (ii) motion saliency map derived from a spherical motion field feature. The resulting dynamic saliency map is computed by integrating both saliency maps.

2.1. Static Saliency Map on the Sphere

The input signal is a color image defined on the sphere, i.e. in spherical coordinates $\theta \in [0, \pi], \varphi \in (0, 2\pi]$ (the radius of the sphere is $r = 1$). Originally proposed in [4], the model is based on four main steps. First, several static features are extracted from the image. Second, for each feature, a multiscale contrast detection method is applied to compute their corresponding conspicuity map. This map highlights the contrasts at different scales according to the feature. Third, the features of the same nature are integrated in order to define a set of cues. Finally, all the cues are integrated into the saliency map. This scalar map indicates salient locations, in the form of a saliency distribution. Precise details on the computation of the static saliency map performed on the sphere can be found in [5].

Let us remind the final integration step. The resulting static spherical saliency map is computed by fusing together all cue conspicuity maps:

$$S_{S^2} = \sum_{cue \in int, chrom, orient} \mathcal{N}(C_{cue}(\theta, \varphi)), \qquad (1)$$

where $\mathcal{N}()$ is the map integration scheme that simulates the competition between the maps. In this paper, we consider three cues: intensity, color and orientation. We note that the conspicuity maps are previously scaled at the same range values by applying a peak-to-peak normalization.

2.2. Motion Saliency Map on the Sphere

The motion saliency map highlights the most salient regions according to the motion feature. The motion model can be described in two steps:

- The estimation of the spherical motion feature. It corresponds to a spherical motion field pyramid, used as input feature of the motion model (Subsection 2.2.1).
- Applying a multiscale motion contrast detection method to the spherical motion feature (Subsection 2.2.2).

2.2.1. Spherical Motion Field Pyramid

The spherical motion field pyramid Π_M (Figure 1) is composed of N multiscale motion fields $\mathbf{M}_i((\theta, \varphi))$ on the sphere, $i \,\epsilon\, \{1, 2, ..., N\}$, corresponding to motion vector estimation at different scales. Coarse scale maps detect motion of large regions while fine scale maps detect motion of small regions. The initial resolution of the first level $\mathbf{M}_1$ (the highest resolution) is $h_1 \times w_1$ and the resolution of the other levels is decreasing over the pyramid by factor of 2 between two consecutive levels.

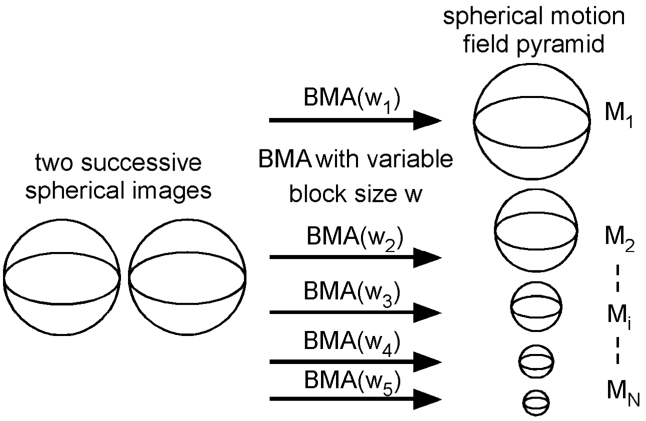
spherical motion field pyramid
two successive spherical images
BMA with variable block size w
BMA(w1)
BMA(w2)
BMA(w3)
BMA(w4)
BMA(w5)
M1
M2
Mi
MN

Fig. 1. Spherical motion field pyramid based on BMA on the sphere.

Each level of the pyramid is estimated using block matching algorithm (BMA) operating directly in spherical coordinates [7]. Basically , the idea is to partition the spherical grayscale image into uniform solid angles of size $B_i \equiv b_{i,\varphi} \times b_{i,\theta}$, called spherical blocks. These blocks are then paired with the best matching blocks of the same size in the reference spherical image within a search window of size $s_{i,\varphi} \times s_{i,\theta}$.

To compute the spherical motion pyramid Π_M, the block size B_i is varying according to the pyramid level $\mathbf{M}_i$ in order to detect large moving regions with large blocks and fine moving regions with fine blocks (Figure 1). The block size is therefore computed according to the following equation:

$$B_i = 2^{(i-1)} \cdot B_1, \qquad (2)$$

where B_1 is the initial block size at the first level.

We must note that non-overlapped partitioning is used for defining the blocks. In addition, a full search technique and minimization of the mean square error (MSE) distance is used for the matching.

2.2.2. *Multiscale Motion Contrast Detection Method*

In the sense of visual attention, center-surround contrast refers to a difference between a center and surround region. It is admitted that such contrast can be modeled by DoG filtering. This method has however the inconvenience of being heavy in terms of computation costs, especially for computing center-surround contrasts at numerous scales. For this reason, regarding the static model and what concerns the sphere, an alternative approach has been proposed in [5] to approximate the multiscale center-surround contrast using spherical image pyramid and cross-scale differences. Regarding the motion feature, we use a similar approach, which is described below.

In order to compute motion contrasts, the idea is basically to define two average motion vectors $\vec{v}_c$ and $\vec{v}_s$ from the motion pyramid Π_M, representing respectively the motion of center and surround regions.

Once the motion average vectors $\vec{v}_c$ and $\vec{v}_s$ have been estimated from the motion pyramid, a motion conspicuity operator is applied in order to detect center-surround contrast. Several operators are possible according to the nature of motion contrasts [3]. In this paper, we consider an operator based on motion contrast in magnitude, which is suitable to highlight salient moving regions of video sequences with fixed background. We note that another operator would be required to highlight both phase and magnitude motion contrasts that occur in the case of moving background.

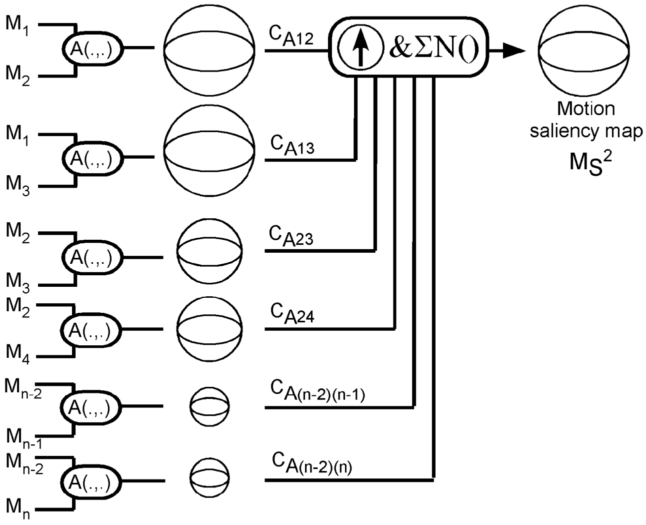

Fig. 2. Motion saliency computation on the sphere.

Formally, the magnitude motion contrast operator computes the norm of the center and surround motion vectors and the absolute difference:

$$A_{cs}(\vec{v}_c, \vec{v}_s) = |\|\vec{v}_c(\theta,\varphi)\| - \|\vec{v}_s(\theta,\varphi)\||, \qquad (3)$$

where $\vec{v}_c$ is the motion vector at the center level $\mathbf{M}_i$ and $\vec{v}_s$ is the motion vector at the surround level which is up-sampled to the corresponding resolution.

The spherical motion pyramid and magnitude operator having been defined, we describe below the multi-scale motion contrast detection method used to compute the resulting spherical motion saliency map.

In order to highlight motion contrast at different scales, several intermediate conspicuity maps are computed, each one corresponding to a specific size of center-surround contrast (Figure 2). The motion magnitude operator A_{cs} is applied several times at the different levels of the pyramid to compute the intermediate conspicuity maps $C_{A_{ij}}$:

$$C_{A_{ij}}(\theta,\varphi) = |\ \|\vec{v}_i(\theta,\varphi)\| - \|\vec{v}_j(\theta,\varphi)\|\ |. \qquad (4)$$

We note that up-sampling is required to perform point-by-point substraction. In this paper, we use a spherical motion pyramid of $n = 6$ levels, each intermediate conspicuity map $C_{A_{ij}}$ is obtained from a center level $i \epsilon \{1,2,3,4\}$ and a surround level $j = i + \delta$ with $\delta \epsilon \{2,3\}$. δ corresponds to the scale difference between the center and surround level. Therefore, 6 center-surround differences are computed at different scales (1-3, 1-4, 2-4, 2-5, 3-5, 3-6). Each intermediate conspicuity map has a resolution corresponding to its center level.

Finally, all intermediate maps are up-sampled at the initial resolution and integrated into the motion saliency map C_{magn} using the same map integration scheme as in the static model:

$$C_{magn} = \sum_{i,j} \mathcal{N}(C_{A_{ij}}(\theta,\varphi)). \qquad (5)$$

2.3. Dynamic Saliency Map on the Sphere and Spot Detection

In this section, we describe the integration of both static S_{S^2} and motion M_{S^2} spherical saliency maps in order to obtain the final dynamic saliency map D_{S^2}. This yields a single saliency measure of interest for each location on the sphere.

The final spherical saliency map is computed according to the following equation

$$D_{S^2} = \mathcal{N}(S_{S^2}) + \mathcal{N}(M_{S^2}), \qquad (6)$$

where $\mathcal{N}(.)$ is the same map integration strategy as in the static model.

Finally, from the spherical dynamic saliency map, the most salient locations on the sphere are selected. The idea consists in detecting successively the locations of the maxima. "Winner-Take-All" (WTA) mechanism and inhibition of return (IOR) are applied iteratively on the saliency map. The complete details can be found in [5].

3. EXPERIMENTAL RESULTS

In order to illustrate the proposed dynamic visual attention model operating on the sphere, we apply it on real omnidirectional image sequences. The video sequence is acquired with a spherical omnidirectional multi-camera sensor (LADYBUG [8]). Each frame from this sequence is defined on 1024×1024 equi-angular spherical grid (θ, φ) and covers around 75% of the sphere. The camera is placed on a table in an office, while a person enters in the room. Two frames from the sequence are shown on Figure 3 (a). The dynamic saliency map is depicted in (b). The resulting spots of attention are determined from the saliency map. The first three of them are shown in (c). We can see that the proposed attentive system highlights both static and moving salient regions in the scene: two spots (#1 and #2) are located on the moving person and the spot #3 corresponds to the salient static object.

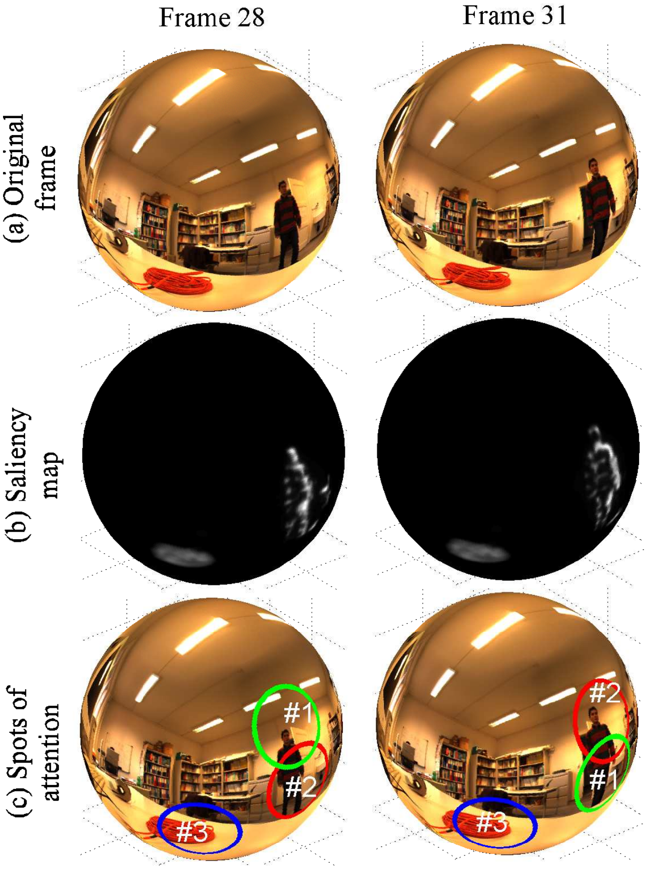

Fig. 3. Experimental results: (a) input omnidirectional image on the sphere; (b) spherical saliency map; (c) spots of attention on the sphere.

4. CONCLUSIONS

In this paper, we have proposed a dynamic attentive system for omnidirectional video sequences operating on the sphere. The process encompasses the multiscale contrast detection of static and motion features, and its integration to define the resulting dynamic saliency map. Such a system can be a prospective solution to speed-up computer vision applications. Indeed, it can be seen as a preprocessing step, which reduces the amount of information that will be processed later by high-level and computer vision tasks. To illustrate the concept, the proposed system has been applied to video sequences acquired with an omnidirectional multi-camera sensor. Each frame is represented in the spherical coordinates. The experiments illustrated the selection process of both static and motion salient regions, represented by a subset of attentional spots.

Regarding the motion contribution, the proposed system highlights motion contrast in magnitude, which is suitable for video sequences with fixed background. As a future work, the attentive system could be extended to omnidirectional video sequences with moving background.

In addition, we note that the proposed system, which operates on the sphere, could be applied also on other catadioptric omnidirectional sensors once they have been mapped on the sphere. Such an extension requires a geometric transformation. Besides, in [9] was proposed a solution to extend a static attentive system on the sphere to hyperbolic and parabolic omnidirectional images.

5. REFERENCES

[1] L. Itti, "Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes," *Visual Cognition*, vol. 12, pp. 1093–1123, 2005.

[2] O. Le Meur, and P. Le Callet, and D. Barba, and D. Thoreau, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, pp. 2483–2498, 2006.

[3] A. Bur, *Computer models of dynamic visual attention*, Ph.D. thesis, Université de Neuchâtel, Switzerland, 2009.

[4] Ch. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[5] I. Bogdanova, A. Bur, and H. Hügli, "Visual attention on the sphere," *IEEE Transactions on Image Processing*, vol. 17, pp. 2000 –2014, 2008.

[6] C. Geyer and K. Daniilidis, "Catadioptric projective geometry," *International Journal of Computer Vision*, vol. 45(3), pp. 223–243, 2001.

[7] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," in *EUSIPCO*, 2005.

[8] LADYBUG, *http://www.ptgrey.com/products/spherical.asp*.

[9] I. Bogdanova, A. Bur, and H. Hügli, "The spherical approach to omnidirectional visual attention," in *EUSIPCO '08: Proceedings of the 16th European Conference on Signal Processing*, Lausanne, Switzerland, 2008.