2011 International Conference on Telecommunication Technology and Applications
Proc .of CSIT vol.5 (2011) © (2011) IACSIT Press, Singapore

Multi Camera Visual Saliency Using Image Stitching

Christopher Wing Hong Ngau[+], Li-Minn Ang, Kah Phooi Seng

School of Electrical and Electronic Engineering
The University of Nottingham
43500 Selangor, Malaysia

Abstract. This paper presents and investigates two models for a multi camera configuration with visual saliency capability. Applications in various imaging fields each have a different set of detection parameters and requirements which would result in the necessity of software changes. The visual saliency capability offered by this multi camera model allows generic detection of conspicuous objects be it human or non-human based on simple low level features. As multiple cameras are used, an image stitching technique is employed to allow combination of Field-of-View (FoV) from different camera captures to provide a panoramic detection field. The stitching technique is also used to complement the visual saliency model in this work. In the first model, image stitching is applied to individual captures to provide a wider FoV, whereby the visual saliency algorithm would able to operate on a wide area. For the second model, visual saliency is applied to individual captures. Then, the maps are recombined based on a set of stitching parameters to reinforced salient features present in objects at the FoV overlap regions. Simulations of the two models are conducted and demonstrated for performance evaluation.

Keywords: image stitching, multiple cameras, panoramic saliency map, visual saliency

1. Introduction

The availability of low cost cameras and wireless technology has resulted in a widespread deployment of such devices in both commercial and private sectors. Cameras are widely used in surveillance, especially in crime prevention and traffic monitoring. Besides the norm, they are also used for quality improvement and statistical purposes such as patient counting at health care centres and redirection of human queues at commercial areas. With a massive deployment of cameras in an enclosed area, constant surveillance can be a tedious task for the monitoring personnel. For real-time surveillance, a person can effectively monitor one to four feeds at a single instance [1]; not to mention several hundreds of live feeds that have to be monitored at areas such as shopping malls and security-imposed locations.

Automated systems are slowly out-phasing the conventional monitoring method; keeping up with the gradually increasing number of camera deployments. Automated systems allows monitoring and detection of objects. In certain cases, classification of the objects according to its actions and features is possible with mathematical and image processing solutions [2]. Recent advances allows facial detection and recognition with active cameras [3]-[4], providing an advantage in terms of security or authorization feature. Such unmanned systems also allow continuous data collection which then can be analyzed at a later time.

With the advantage of widely deployed automated systems, there is a main drawback: tremendous amount of incoming visual information that is required to be processed or stored. Although the massive stream of visual information is a trivial issue for systems solely made for monitoring purposes, it can be an issue for advanced systems which performs high level processing in real-time. To effectively process a large amount of visual information satisfactory in real-time requires additional hardware resources which proves to

[+] Corresponding author. Tel.: + 6(03) 8924 8622; fax: +6(03) 8924 8017.
E-mail address: Wing-Hong.Ngau@nottingham.edu.my.

be cost inefficient. Also, with the increased number in data collection, the required amount of storage medium increases significantly if no data compressions are performed [5].

In order to allow satisfactory processing of information without incurring additional hardware cost, it is often desirable to remove redundancy in the data before any high level processing takes place. However, the data filtering itself has a problem as different application may have its own set of selection criteria and information filtering which is usually task specific. This would make the filtering process more difficult if the camera system is used for several tasks. The solution to obtain a generic information filtering may lie in how early information is processed in the human visual system. It is believed that the visual system in humans consists a set of visual filters [6] which extract relevant information based on low-level features such as luminance, colour, edges, etc [7]; where these features serve as stimuli in providing objects the perceptual quality of being important or unique.

The perceptual characteristic of an object that makes that object stand out from its surrounding is described with the term visual saliency; in most cases, causing the redirection of attention towards it. The term visual saliency is often used interchangeably with visual attention [8]-[9]; describing importance, uniqueness, or rarity of an object or region. There are various research works regarding the area of visual saliency [10]-[12] with the aim of reproducing the function of the human visual system that will compliment and aid processing techniques relating to imagery and machine vision. With most visual saliency models, there will be an output saliency map in which its topological amplitudes determine how conspicuous objects and regions are in the actual visual capture. The saliency map is generated from parallel processing of extracted low-level features exhibited in the visual capture alone. Here, the visual saliency can be considered a bottom-up low level processing [13] which filters out irrelevant information prior to high level processing which are task-dependent.

The number of cameras deployed at a location usually consists of several units, each placed at a strategic location. In such cases, such a number is required as each unit has a limited Field-of-View (FoV) [14]. By having multiple cameras at several locations, the inability to cover blind spots which are out of a single camera's FoV can be compensated. It is possible to widen the FoV of a single unit through the use of fish-eye lenses; however, distortion usually occurs with this method. In the presented models, an image stitching technique is used to combine multiple FoVs into a single panoramic FoV with minimal distortion. With the stitching method, the large amount of image feeds can be reduced.

In this paper, two models for the multi camera using visual saliency are presented and demonstrated through simulations. In the first model, images from multiple cameras are stitched and visual saliency is applied to the stitched image. This method allows a wide coverage from several angles or views. In the second model, information from multiple cameras are used to further pin-point conspicuous objects through saliency map reinforcement. This is particularly useful in reinforcing the features from objects appearing in overlap regions of the camera's FoV as illustrated in Fig. 1.

The paper is organized as follows: Section 2 covers the equations of the visual saliency algorithm used in the multi camera model. Section 3 gives a brief account on the image stitching technique. Sections 2 and 3 consist of techniques used in both the multi camera models presented in this paper. Section 4 illustrates on the two multi camera models and describes their methods. Section 5 shows the simulation results along model evaluation. Finally, the paper is concluded in Section 6.

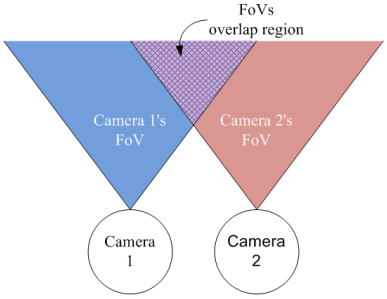

Fig. 1: FoVs overlap region of two cameras.

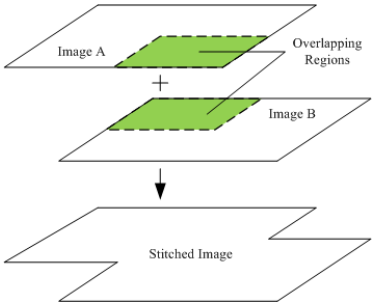

Fig. 2: Stitching of two images.

2. Visual Saliency Algorithm

Visual saliency models can generally be categorized into two main types: biologically plausible and purely computational [12]. The prior models based their methods on the architecture of biological visual system, taking into consideration the mechanism of one or more stages in the visual system. The latter are usually based on mathematical techniques such as statistics and probability or image processing techniques. Many purely computational models have lower computation complexity and work reasonably well with simple visual scenes. However, that is not the case with complex scenes, which is what cameras nowadays capture. In this paper, a biologically plausible visual saliency model which operates on a range of medium spatial frequencies [15] is used. Recent studies showed that human gazes on conspicuous objects peaked at a range of medium spatial frequencies [6] [15]. The algorithm is described in the following paragraphs.

The input image is first converted from the RGB color space to the YCbCr color space. The next step is to extract early visual features from the input image. The 9/7 Cohen-Daubechies-Feauveau (CDF) wavelet transform is applied to each individual Y, Cb, and Cr channels of the image. For each channel, a five-level decomposition is applied. The CDF wavelet transform separates the image at each level into four frequency sub-bands: LL, HL, LH, and HH. As medium spatial frequencies are of interest, only the orientation sub-bands: HL, LH, and HH are considered.

Once the five-level wavelet pyramids for each channel is obtained, a centre-surround (CS) process is applied to each orientation sub-band to obtain orientation maps. For each location, the CS filter is computed as the difference between the current pixel coefficient value and the average coefficient value of the surrounding pixels using Eq. 1:

$$CS(x) = \left| |I(x)| - \frac{1}{s} \sum_{k \in S} |I(k)| \right| \tag{1}$$

where $I(x)$ is the pixel coefficient value at location x, S is the surround support, and s is the surround area. The surround area is fixed to 5×5 pixels for each pyramid level.

The next step consists of summing up the orientation maps according to pyramid level to form the feature maps. Then for each level, a filtering process is applied as shown in Eq. 2:

$$L(x) = \frac{1}{d} \sum_{k \in D(x)} \left(\sum_{o \in (1,2,3)} CS_o(k) \right) \tag{2}$$

In Eq. 2, $CS_o(k)$ is the center-surround response at location k for the orientation sub-bands, $D(x)$ is disk of radius 1° centered on k, and d is the surface area of the disk.

The fusion stage of the algorithm consist of map combination. In this step, features maps of luminance (from Y channel) and color (from Cb and Cr channels) are fused together according to channel by means of bilinear up-sampling and point-to-point addition shown in Eq. 3:

$$C(x) = \frac{1}{lv} \bigoplus_{l} L_l(x) \tag{3}$$

where lv is the number of decomposition level and $L_l(x)$ is the level l map.

The output saliency map is finally obtained by point-to-point addition of the fused maps from the three channels and normalized to the range between 0 and 255 (8-bit greyscale).

3. Image Stitching For Multiple Cameras

Image stitching is a technique used to join multiple images (with overlapping regions) from different cameras into a same coordinate system. This technique can be used to extend the effective FoV as shown in Fig. 2. In cases where an object is partially captured by two separate cameras, the object can be recombined using this technique for better recognition. In addition, objects that are obstructed in a camera's FoV can be captured in another camera's FoV; and are eventually placed on a single plane system for more efficient

analysis [16]. These advantages seen in the stitching technique is favourable in applications which makes use of object detection and recognition.

To stitch two or more images which are not aligned on a same coordinate plane due to camera placements, feature points have to be extracted to allow matching between these images. Due to scale and rotational differences in images, the Scale-Invariant Feature Transform (SIFT) [17] is used to extract and match feature points across the images. Consider a case of two images captured by two separate cameras. The second image can be seen as a geometric transformation of the first image where the transformation matrix can be written as:

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} \approx \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

Expanding the equation and rearranging would result in two linear equations per matching feature:

$$\begin{aligned} x' &= \frac{a_1 x + a_2 y + a_3}{c_1 x + c_2 y + c_3} \\ y' &= \frac{b_1 x + b_2 y + b_3}{c_1 x + c_2 y + c_3} \end{aligned} \tag{5}$$

After the matching features have been found, a K-D tree is used to determine the nearest neighbour for each matching pair of feature points.

4. Multi Camera Visual Saliency Models

In this section, two models are presented for the multi camera with visual saliency capabilities. In the first model shown in Fig. 3(a), the visual saliency is applied to the stitched image. This model covers a wider FoV and can be advantageous especially when the object of interest is constantly appearing across individual camera's FoV. In cases where an object is partially captured between two FoV, the stitching technique allows the object to be represented as a whole, making detection much easier.

Whereas for the second model shown in Fig. 3(b), the visual saliency is first applied individually to the two captured image. Then, stitching is performed to generate a set of stitching parameters to aid the process of saliency map fusion. The information from the two saliency maps are fused together, providing a complete map. This allows reinforcement of salient features of objects appearing in the overlapping FoV; hence improving the accuracy of detection in the high level processing.

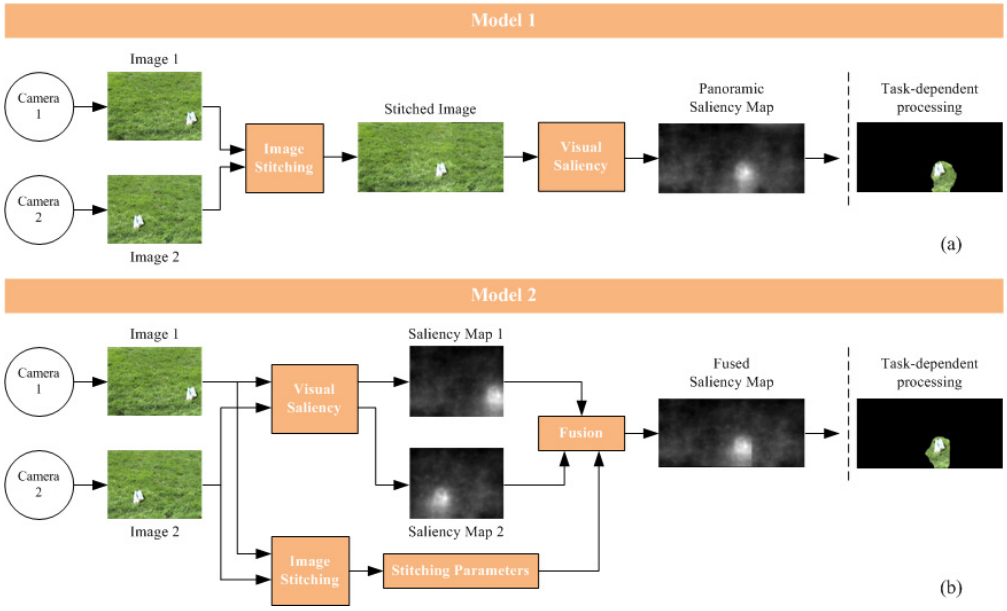

Fig. 3: Two multi camera visual saliency models of interest. Test image is taken from [11].

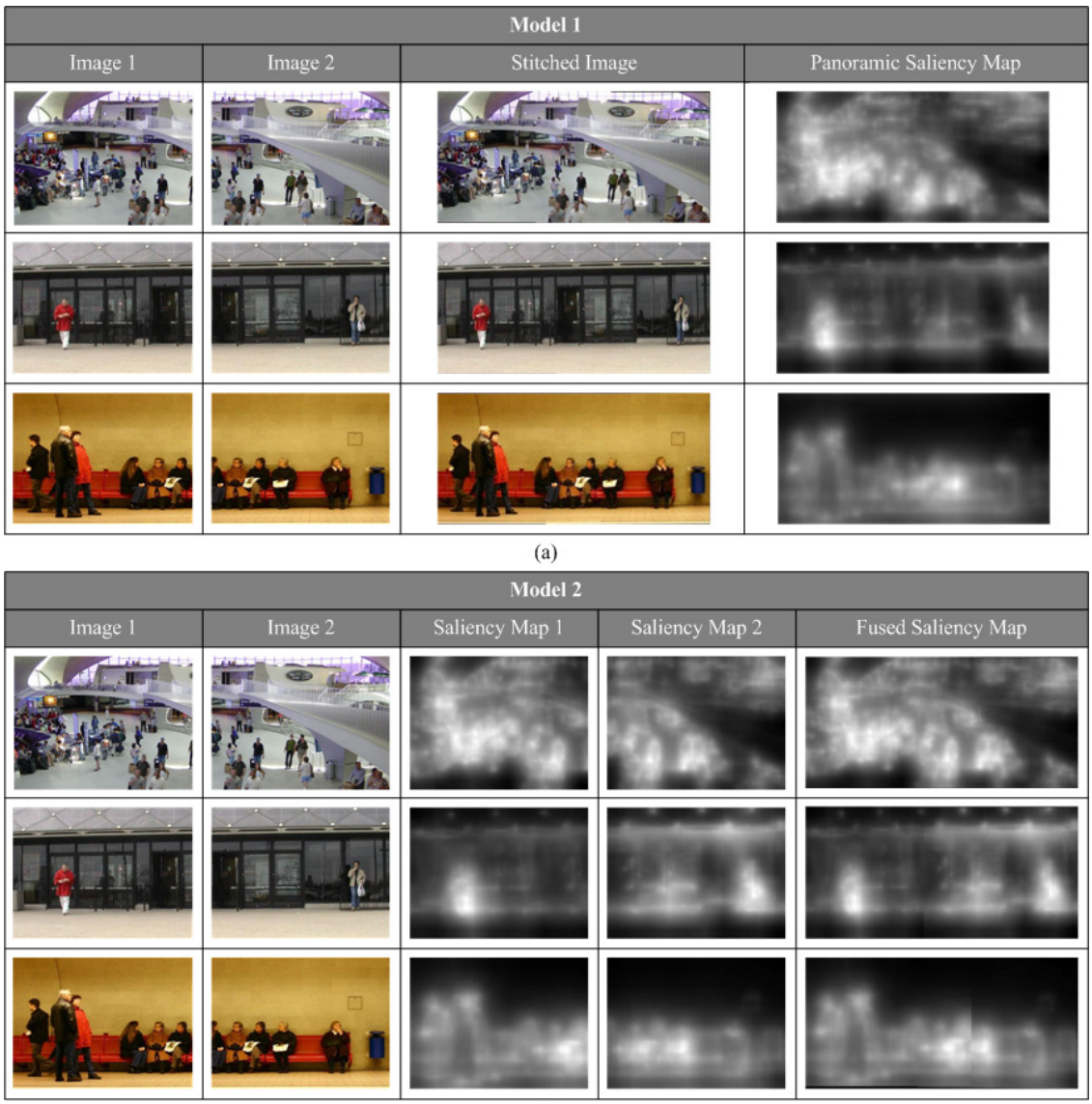

Fig. 4: Simulation results for (a) Model 1 and (b) Model 2.

5. Simulation Results and Evaluation

The two multi camera models are simulated using complex scene images from the MIT Indoor Scene Recognition Database [18]. The results for both models are shown in Fig. 4(a) and Fig. 4(b) respectively. Both models only covers task independent low level processing whereas high level task dependent processing can be varied accordingly with desired applications.

5.1. Discussion

In the first model (Fig. 4(a)), stitching is performed on the two images before visual saliency is applied. As multiple matching feature points are used, the stitched image appears to be seamless although some distortion may occur at the borders. With the stitched image, the visual saliency algorithm is able to provide a smooth gradient saliency map spanning over a wide area. Model 1 can be seen as a method with a panoramic saliency map which allows detection of conspicuous objects and regions over a wide FoV.

For the second model (Fig. 4(b)), visual saliency is applied to individual image to obtain multiple saliency maps. At the same time, feature matching is performed on the captured images to extract matching points which will help the fusion of the saliency maps. As conspicuous feature detection is performed on separate images (can be considered as local operation), the features are enhanced during the fusion stage. Therefore, Model 2 is seen to be a method used to enhance local conspicuous features in multiple images (local) in comparison to Model 1 (global).

5.2. Model Evaluation

From comparisons of results in Figure 4, it can be seen that the output map for Model 1 appears to be seamless and artefact-free whereas distortion occurs for Model 2. This is due to the information losses at the image borders before the fusion stage for Model 2. Regardless of this drawback, Model 2 provides a stronger prediction of salience than Model 1. Local salient feature processing emphasizes on conspicuous objects in a smaller neighbourhood rather than in a wide area as in Model 1. This causes strong features to be "reinforced" when the saliency maps are fused. Depending on the nature of application, either Model 1 or Model 2 can be seen as favourable.

6. Conclusion

This paper has presented two models for multi camera; both generating an output map with the aid of visual saliency and image stitching, which will serve useful to high level task dependent applications. The works in this paper also demonstrated a panoramic saliency map which encodes conspicuous features from a wide FoV and a fused saliency map which reinforces conspicuous features in multiple image captures. For future works, the authors intend to investigate the capabilities of the presented models in solving occlusion and lighting problems using live test images.

7. References

[1] H. Dee and S. A. Velastin. How close are we to solving the problem of automated surveillance? A review of real-world surveillance. *Machine Vision and Applications*. 2008, **19**(5-6): 329-343.

[2] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In: *Proc. of the 7th European Conference on Computer Vision*. 2002.

[3] P. Petrov, O. Boumbarov, and K. Muratovski. Face detection and tracking with an active camera. In: *Proc. of the 4th IEEE International Conference on Intelligent Systems*. 2008, pp. 14-34.

[4] Z. Zhang, G. Potamianos, M. Liu, and T. Huang. Robust multi-view multi-camera face detection inside smart rooms using spatio-temporal dynamic processing. In: *Proc. of the 7th IEEE International Conference on Automatic Face and Gesture Recognition*. 2006, pp. 407-412.

[5] B. A. Wandell. *Foundations of Vision*. Stanford University: Sinauer Associates Inc., 1995, pp. 249-250.

[6] G. Mather. *Foundations of Perception*. Psychological Press, 2006.

[7] J. Wolfe. *Visual Search in Attention*. University College London, 1998, pp. 13-74.

[8] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In: *ACM International Conference on Multimedia*. 2003.

[9] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*. 1995, **78**(1-2): 507-545.

[10] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shift of visual attention. *Journal of Vision Research*. 2000, **40**(10-12): 1489-1506.

[11] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. *Advances in Neural Information Processing Systems*. 2006, pp. 155-162.

[12] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In: *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1597-1604.

[13] A. Benoit, A. Caplier, B. Durette, and J. Herault. Using human visual system modeling for bio-inspired low level image processing. *Computer Vision and Image Understanding*. 2010, **114**(7): 758-773.

[14] P. Pstiak. Implementation of HDR panorama stitching algorithm. In: *Proceedings of the Central European Seminar on Computer Graphics*. 2006.

[15] F. Urban, B. Follet, C. Chamaret, O. L. Meur, and T. Baccino. Medium spatial frequencies, a strong predictor of salience. *Cognitive Computation, Biomedical and Life Sciences*. Springer. 2010.

[16] W. C. Chia, L.-M. Ang, and K. P. Seng. Performance evaluation of feature detection in using subsampled images for image stitching. In: *Proc. of the IEEE International Conference on Computer Science and Information Technology*. 2010, pp. 60-64.

[17] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*. 2004, **60**(2): 91-100.

[18] Indoor Scene Recognition Database:
http://web.mit.edu/torralba/www/indoor.html.