Conference Paper

Image matching principles in photometrical transformations

Cunningham, S. and Suprun, D.

This is a paper presented at the 7th IEEE Int. Conference on Internet Technologies and Applications ITA-17, Wrexham, UK, 12-15 September 2017

Image Matching Principles in Photometrical Transformations

Diana Suprun
Bauman Moscow State Technical University
Moscow, Russia
Dianasuprun91@gmail.com

Stuart Cunningham
Glyndŵr University
Wrexham, Wales
s.cunningham@glyndwr.ac.uk

***Abstract*—In this article interactive visualization techniques for creating virtual objects are considered. We describe the most common methods of photometric transformations between images and a variety of geometric objects contextualized against the backdrop of increased adoption of 360 videos and virtual reality systems. Two techniques, the Harris-Laplacian and the Scale Invariant Feature Transform (SIFT) have been described. The algorithm estimation of virtual objects interactive visualization is given. The image-matching algorithm using key points is described.**

Keywords — three-dimensional object; Harris-Laplacian algorithm; Scale Invariant Feature Transform; Gaussian pyramid; Difference of Gaussian pyramid; key point; image matching algorithm.

I. INTRODUCTION

In recent years, there has been a significant and renewed interest in developing interactive software that allows a user to explore a virtual reality (VR) or augmented reality (AR) environments. In part, this has been due to the revival of stereoscopic displays, such as the Oculus Rift, Google Cardboard, and Samsung Gear devices. These are being applied in a huge variety of application scenarios such as the visualization of big data [1]; in medical scenarios, such as management of pain [2] and biomechanics [3]; new modes of interaction [4] and experiences in games and entertainment [5]; in enhancing education and pedagogy [6, 7]; and as potential assistive technology for navigation tasks [8, 9].

The presentation of these virtual elements is created using a variety of techniques, such as multi-camera video recordings, computer generated images, or using computer game engines, such as Unity or Unreal. However, this predominantly means that all such systems utilize 360 degree imaging and 3D graphics. From a software and hardware perspective, producing these images in near real-time, and being able to re-render scenes rapidly in response to user movement and interaction, is not a trivial task. As such, this work of this paper is concerned with exploring mechanisms by which the rendering of these graphics might be improved, whilst remaining robust to the transforms likely to be required by user movement and interaction in the virtual space.

Consider the challenges that occur during the process of capturing photographs. One of these challenges is to use natural lighting. In the case of an indoor photograph, for example, the area of a nearby window may be overexposed, making it difficult to deal with image stitching as two adjacent pictures may different light characteristics. To solve this the white balance can be controlled manually resulting in the value of the white balance being different in every picture. In addition, at the time of a panorama snapshot there can be a change of exposure. To deal with this a larger overlap between adjacent frames is required. Respectively, the number of pictures for the closure of the virtual panorama scope is increased. Whilst discussing this optimization process it can be mentioned that it in our own work of developing a virtual tour it was necessary to get a three-dimensional panorama by matching together key points using 26 images. The main challenge was to deal with discrepancy of the camera field of view. That is why the optimization by means of the control point's distortion was required.

The considered method is to derive distinctive invariant features from images to make effective matching between various views of scene or an object. These features are invariant to rotation and image scale, and are to provide effective matching in different cases, such as: change in 3D viewpoint or illumination, addition of noise and so forth. The process of recognition consists of matching individual features to a database of features from known objects [10]. This approach is to distinctly identify objects in situations where occlusion and clutter may occur whilst maintaining near real-time performance. For the computer, the image is a set of data points, so to accomplish image matching appropriate methods should be used. There are methods of image matching based on a comparison of knowledge about the images. [11]. Their idea is to calculate the value of a particular function for each point of the image. Based on these values certain characteristics of the image can be assigned and then the problem of image matching is reduced to the comparison of such characteristics. The advantage of this method is its simplicity; the disadvantage is that they work only in ideal situations. In the case of noise being present in the images, or changing of scale, the application of such techniques becomes meaningless. This is explained by the fact that each point on the image contributes to its characteristic. To avoid this problem, the entire set of points to allocate special or key points is needed, and only then

978-1-5090-4815-1/17/$31.00 ©2017 IEEE

should they be compared. It became the foundation of another embodiment method of image matching using key points.

There are two appropriate techniques for this purpose: the Harris-Laplacian (HL) technique and the SIFT (Scale Invariant Feature Transform) algorithm.

II. The Harris-Laplacian technique

The HL technique is used to compare images irrespective of photometric transformations or geometric transformations between images and to identify appropriate or corresponding areas between the images. The advantage of the technique is that it allows finding features that are robust to illumination changes and invariant to scaling and image rotation. A description is designed for each feature by means of the local neighbourhood and then it proceeds as a unique identifier for the feature [12]. These identifiers are used to recognize "point to point correspondences" between images [13]. This method can be used to execute image retrieval for panoramic images. Stages of execution of the technique for three-dimensional reality are as follows:

- Computation of the image gradient at each point using Gaussian smoothing;
- Calculation of the *M* matrix with the weights for the Gauss window neighborhood of each point;
- Calculation of the *R* angle response measure;
- Point detection with a large value of *R* (clipping threshold);
- Detection of the local maximum of the corner response measures (non-maximum suppression) [14].

The main reason for the widespread use of the HL technique is its invariance to rotation (Fig.1) and intensity shift. This method allows developers to work more effectively in terms of image quality [14].

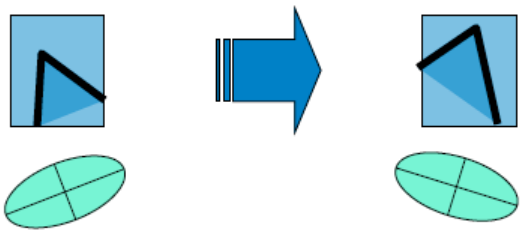

Figure 1. HL technic invariant to rotation [15]

As for the criticism, the method has a drawback because it is not invariant to scale (Fig.2) [14]. In VR and AR situations, users will often be able to move around amongst the virtual objects being presented, in addition to being able to rotate their field of view. To this extent, rotation alone may not be the only type of affine transform required, this is why the SIFT algorithm should be used.

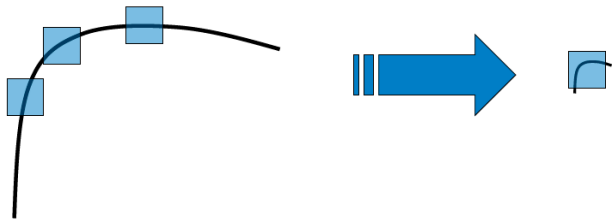

Figure 2. HL technic isn`t invariant to scale [15]

III. The Scale Invariant Feature Transform method

The SIFT method is now widely used in image matching process of interactive visualization. The method works as follows: for each point in the image the value of a particular function is computed. Based on these values certain characteristics of the image can be assigned. Thus, the process of image matching is down to the comparison of these characteristics. The advantage of the SIFT method is its simplicity. The disadvantage of this method is that it works mostly in ideal situations. According to research on the subject [16], it may be caused by several factors, such as: appearance of new objects on the image, overlap of one object to another, noise, zoom, position of the object in the image, a camera position in three-dimensional space, lighting, affine transformations, etc. It can be explained by the fact that each point on the image contributes to the characteristic. Thus, to compare these characteristics the entire set of key points is needed. We can conclude that there is a need to solve this limitation. The task is to somehow choose points that contribute to the characterization, or, even better, to allocate some special (key) points and compare them. It brings to the idea to compare images using key points [17]. The image is replaced by a model; a set of key points. It should be noted that the key point is a point on the image of an object that is very likely to be found in another image of the same object. The detector is the method of extracting key points from the image. The detector will provide resistance to the problems of scaling [17]. Consequently, the invariance to scale issue can be solved.

The development process is divided into iterations. The main steps of SIFT method [17] are as follows:

- Scale-space peak selection;
- Difference-of-Gaussians (DoG) implementation;
- Key point localization;
- Elimination of unstable key points;
- Orientation assignment based on key point local image patch;
- Key point descriptor selection based upon the image gradients in key point local neighborhoods [17].

Taken together, the above-mentioned features of SIFT descriptors, it should be noted that this technology has some drawbacks. Not every point and its description will meet the requirements. This can have unwanted consequences upon future process of image matching. In some cases, the solution may not be found even if it exists. For example, consider the situation when two images have representations of a brick wall.

The solution can`t be found due to the fact that the wall is composed of repeating objects (bricks). So, for different control points, the descriptions are similar. Despite this limitation, descriptors usually work well in many cases of practical importance.

The SIFT method is used to find key points on the image based on the Gaussian pyramid. To identify key points on the image the Difference of Gaussian pyramid (DoG) is designed [11].

The Gaussian is an image that is flooded by a Gaussian filter. It is described as (1):

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

Where $L(x, y, \sigma)$ is a Gaussian value at the point (x,y) and σ represents a blur radius; $G(x, y, \sigma)$ is a Gaussian kernel; $I(x, y)$ is an original image value [11].

The DoG is a resultant image formed by means of per-pixel subtraction of the Gaussian original picture and the picture with a different Gaussian kernel. The DoG is expressed as:

$$D(x, y, \sigma) = [G(x, y, k\sigma) - G(x, y, \sigma)] * I(x, y) = L(x, y, k\sigma) - G(x, y, \sigma) \quad (2)$$

As a result, a scalable image space is a variety of different, smoothed options of the original image. Compared to the Harris-Laplacian space the Gaussian scalable space is linear, shift-invariant, and invariant under rotation and scale.

According to Lowe [12], invariance under scale is achieved by finding the key points of the original image, taken at different scales. Therefore, a pyramid of Gaussians and Difference of Gaussian pyramid is built (Fig.3).

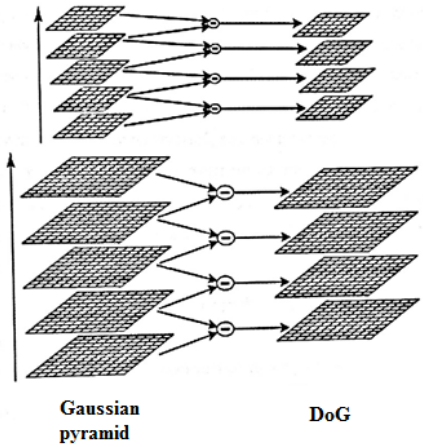

Figure 3. Gaussian pyramid and DoG [12]

Fig. 3 is a schematic representation of DoG. We can see number of differences is one less than the number of Gaussians, and why after transition to the next octave the image size is halved.

After the construction of the pyramids, it is necessary to determine a key point. The point is the key point, if it is a local extremum of difference of Gaussians.

The local extreme point on each image of the DoG is searched. Each point of the current image of DoG is compared with its eight neighbors and with nine neighbors in the DoG that is on the level above and below the pyramid. If a point is greater than, or less than, all neighbor points then it is taken as a point of local extreme. This explains why two additional images in the octave were required.

Once we have found the point of extremum, it is necessary to verify whether they are suitable as key points. At this point, it is necessary to approximate the function of DoG matrix by Taylor polynomial of the second order taken at the point of extremum.

IV. The Image-Matching Algorithm Using Key Points

Consider a method to derive distinctive invariant features from images to make effective matching between various views of scenes or an object, such as would be encountered in a virtual tour of 360-degree visualization. These features are invariant to rotation and image scale, and are to provide effective matching in different cases: change in 3D viewpoint or illumination, addition of noise and others. The image-matching algorithm using key points is presented in Fig.4.

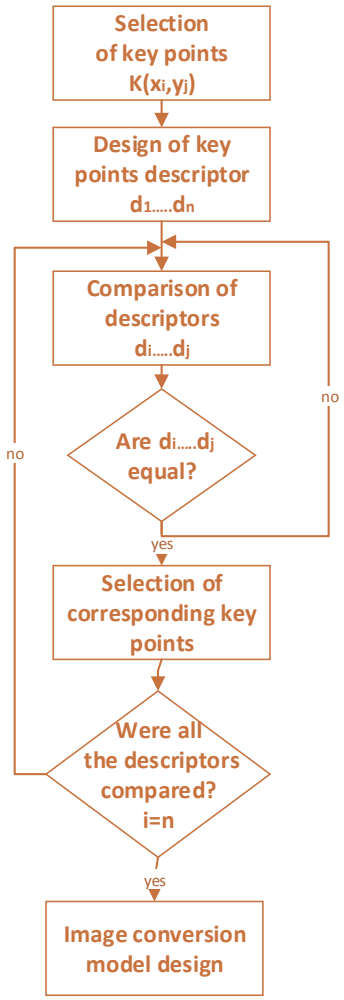

Figure 4. The image-matching algorithm

In the case of image stitching to create panoramic composites, it is necessary to convert the original photos to a format suitable for crosslinking (cylindrical or spherical projections). Crosslinking is the combination of the equal elements located at the common areas of adjacent images. This necessitates mixing images with a view to aligning their brightness, contrast and color tone, ensuring consistency of presentation.

At the stage of photo processing the main features of a panorama's quality are:

- Ability to handle geometric mismatch of adjacent images and the ability to virtually eliminate the phantom images of the elements (for example when there is a fragment of an unnecessary object in one picture, but it is not on the adjacent picture);
- The ability to cope with the color tone difference between adjacent frames (error of mixing sections with color gradient nature (ceiling)).

Thus, the layered file in the Adobe Photoshop (*.PSD) format is designed. It consists of masked transformed images displaced relative to each other and forming the panoramic composite.

V. IMPLEMENTATION EXAMPLE

Consider a simple implementation example of the image-matching approach in the case of creating a 3D panorama. The process consists of several steps. To start, all the photos of a single virtual panorama should be downloaded, in this case to the Adobe Flash environment. The image matching process can be performed automatically in Adobe Flash. However, the resulting image will be full of stitching errors. That is why, to compare overlapping photos control points are generated (Fig.5).

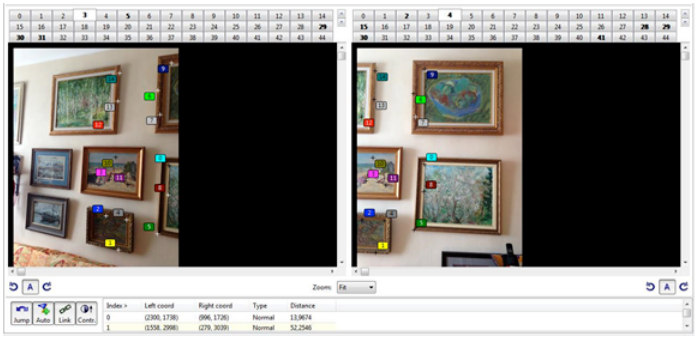

Figure 5. Image matching process

As shown in Fig.5, the program is trying to find identical points in both images. These points are generally angle points. The intensity of an angle point varies relative to the center. Thus, the coordinates, and changes in brightness of the surrounding image points, define the angles points. The main property of these points is distinguishability. This means that there are two dominant gradient directions in the area around the corner. Gradient is a quantity vector that shows the steepest increase in the intensity of the image function $I(x, y)$. To make a more precise matching, we use the top panel with key point numbers to move items. In addition, there is a possibility to correct Gaussians function coordinates that are available in the configuration panel.

Pan (Min/Max/Def): 0.0/ 360.0/ 0.0

Tilt (Min/Max/Def): -90.0/ 90.0/ 0.0

FoV (Min/Max/Def): 10.0/ 120.0/ 70.0

As a result, we obtain a preview of the conformal projection. Substantial image editing was not required in this case because a panoramic head was used when taking the photographs, with a properly set nodal point.

The next step is optimization. Optimization is necessary for the selection of the best lens distortion transformation parameters. Before this step, there are 26 images with distant key points (Fig.5). The objective of the optimization process is to get a three-dimensional panorama by matching together these images' key points. The main idea of the distortion is to deal with discrepancies in the camera's field of view. One object is generally captured twice and by means of this object Adobe Flash is stitching every two images. As the camera is rotating one object is capturing from a different field of view. This is why the optimization, by means of distortion, is required.

Even after automatic stitching and optimization the manual binding of individual elements in the panorama key points is required. This procedure is carried out for every virtual panorama. When gluing, it is recommended that conformal projection panoramas are kept in JPEG format. This will provide further convenience in working with graphic editors. It is worth mentioning that there were some difficult aspects. Since equidistant projection does not preserve parallel lines, then the image is inconvenient to adjust. In order to get the most realistic spherical panorama, it is necessary to remove every distortion from the resulting image, shades and the effect of "vagueness".

VI. APPROACH EVALUATION

To test the usability of the Imaging System that implements the proposed approach, we adopted a method of questionnaire, based upon participants' responses to the images produced by the system.

Ten users were involved in the experiment. The procedure of obtaining decisions involves participants expressing their views on usability issues, such as navigation, ease of use, realistic issues, quickness in response, and others. The method of scores was used to provide a detailed data of expert opinions and helped to analyze the results. To provide a high validity of results different participants from various fields took part in the research. There were five students, two web designers, and three Internet users. The goal of the research was to analyze whether the system is effective in terms of usability issues. The participants were asked to complete a questionnaire with scores for each question in a table. A 1 to 5 Likert scale was used, for participants to indicate the worst to the most pleasant in terms of use.

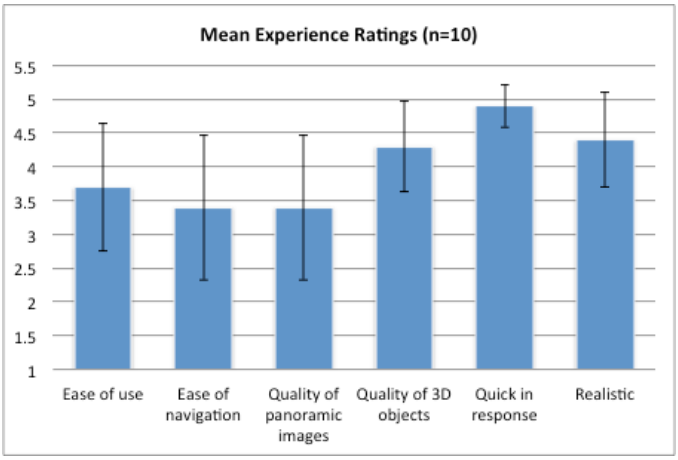

Figure 6. Evaluation of Product

Highest rated is "*Quick in response*" (mean = 4.9; SD = 0.32), whilst lowest, jointly are "*Ease of Navigation*" (mean = 3.4; SD = 1.07) and "*Quality of Panoramic Images*" (mean = 3.4; SD = 1.07). Participants are broadly satisfied with the Imaging System, the image matching approach implementation and their assessment indicates that the System has merit in being adopted for further usage.

VII. Conclusion

This article considered an image-stitching algorithm for three-dimensional interfaces. The results provided have shown a good usability experience. Two methods of image matching were used. It was decided to combine the practicalities of Harris-Laplacian method with the versatility of the SIFT algorithm. For each point on the image the function value was calculated. Based on these values the certain characteristics to the image were ascribed. Thus, the problem of image matching moved to comparison of the characteristics.

In this case, key point identifiers - descriptors, were used. Adobe Flash provides the ability to calculate the point of local extremum for a Difference of Gaussians pyramid to determine whether a point is a key point. This combination allows for calculation of local features to define singular points, allocate special pieces invariant to scale, construct feature vectors, and compare local descriptor pairs of images.

The drawback of this algorithm is that manual annotation is needed. This can be performed automatically using the SIFT algorithm. However, professional equipment is needed, such as a tripod with an automatically rotating head, a camera that allows specific zooming, studio lighting, and so on. As the camera is rotating without a professional tripod, one object is capturing from a different field of view. To deal with discrepancies of the camera's field of view the optimization by means of the above-mentioned algorithms is required.

After comparison of the Harris-Laplacian and SIFT methods, it becomes evident that none of the methods will be efficient if used in their original form. This is why the best solution is a combination of the practicalities of the Harris-Laplacian method, whilst obtaining the versatility of SIFT.

References

[1] C. Donalek, S.G. Djorgovski, A. Cioc, A. Wang, J. Zhang. E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake, and S. Davidoff. "Immersive and collaborative data visualization using virtual reality platforms," *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2014, pp. 609-614.

[2] H. G, M. J, RamirezMaribel, S. J, S. R, and P. R, "Feasibility of articulated arm mounted Oculus rift virtual reality goggles for Adjunctive pain control during occupational therapy in pediatric burn patients," *Cyberpsychology, Behavior, and Social Networking*, Jun. 2014. [Online]. Available: http://dx.doi.org/10.1089/cyber.2014.0058.

[3] X. Xu, K. B. Chen, J.-H. Lin, and R. G. Radwin, "The accuracy of the Oculus rift virtual reality head-mounted display during cervical spine mobility measurement," *Journal of Biomechanics*, vol. 48, no. 4, pp. 721–724, Feb. 2015.

[4] S. Yoo and C. Parker, "Controller-less interaction methods for Google cardboard," *Proceedings of the 3rd ACM Symposium on Spatial User Interaction - SUI '15*, 2015.

[5] I. Hupont, J. Gracia, L. Sanagustín and M. A. Gracia, "How do new visual immersive systems influence gaming QoE? A use case of serious gaming with Oculus Rift," *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, Pylos-Nestoras, 2015, pp. 1-6.

[6] P. T. Kovacs, N. Murray, G. Rozinaj, Y. Sulema, and R. Rybarova, "Application of immersive technologies for education: State of the art," *2015 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL)*, Nov. 2015.

[7] H. M. Knight, P. R. Gajendragadkar, and A. Bokhari, "Wearable technology: Using Google glass as a teaching tool," *Case Reports*, vol. 2015, no. apr22 2, pp. bcr2014208768–bcr2014208768, May 2015.

[8] B. D. Sawyer, V. S. Finomore, A. A. Calvo, and P. A. Hancock, "Google glass: A driver distraction cause or cure?," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 56, no. 7, pp. 1307–1321, Oct. 2014.

[9] U. Rehman and S. Cao, "Augmented reality-based indoor navigation using Google glass as a Wearable head-mounted display," *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2015.

[10] R. Kadobayashi, K. Tanaka, "3D viewpoint-based photo search and information browsing", SIGIR '05, pp. 621-622, New York, 2005.

[11] M. Jay, "Do pictures and Virtual Tours really matter?", Interactive Virtual Media, 2014.

[12] D. Lowe, "Distinctive Image Features from Scale-Invariant Key points", Computer Science Department University of British Columbia Vancouver, B.C., Canada, 2004.

[13] F. Schaffalitzky, A. Zisserman, "Multi-view matching for unordered image sets", In European Conference on Computer Vision, Copenhagen, pp. 414-431, Denmark, 2002.

[14] A. Bhatia, "Hessian-Laplace Feature Detector and Haar Descriptor for Image Matching", For the M.A.Sc. degree in Electrical and Computer Engineering, Ottawa, Canada, 2007.

[15] A. Berg, CSE 391/591: Computational Photography and Introduction to Computer Vision (2008). *Local Feature Detection*. Stony Brook University. [online] Available at: http://alumni.media.mit.edu/~maov/classes/comp_photo_vision08f/lect/18_feature_detectors.pdf [Accessed 10 Jul. 2017].

[16] M. Brown, S. Winder, and R. Szeliski, "Multi-Image Matching using Multi-Scale Oriented Patches", Conference on Computer Vision and Pattern Recognition, 2005, pp. 510–517.

[17] Y. Meng, "Implementing the Scale Invariant Feature Transform (SIFT) Method", Computer Science Department University of British Columbia Vancouver, B.C., Canada, 2000.