Journal of Software Engineering

ISSN 1819-4311

Academic Journals Inc.

www.academicjournals.com

Journal of Software Engineering 9 (2): 401-409, 2015
ISSN 1819-4311 / DOI: 10.3923/jse.2015.401.409

An Improved Algorithm for Extracting Video Frame Interest Point

[1]J. Guo Lv, [1]W. Zhe Kong and [2]D. Yue Li
[1]Modern Urban Geomatic Laboratory of the National Mapping Bureau, Institute of Geomatics, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China
[2]School of Earth Sciences, The Ohio State University, 255 S. Oval Mall, Columbus, OH, 43210, United States of America

Corresponding Author: J. Guo Lv, Modern Urban Geomatic Laboratory of the National Mapping Bureau, Institute of Geomatics, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China

ABSTRACT

In order to solve the problem of image matching, the points matching in the fields of digital photogrammetry and computer vision is researched very deeply. The accuracy of the feature points plays an important role in making right matching results. In a large number of the research of video feature extraction algorithm, the study introduces a method of 3-D Harris detector. An algorithm of Gaussian smoothing is applied for eliminating sudden gray change. Corners of all pixels are detected in a video using Harris. We analyse the effect of detection of different objects using the result of strength corner image. The results of experiments show that the 3-D Harris detector is accurate and efficient.

Key words: Harris detector, 3-D, image smoothing, strength

INTRODUCTION

Harris detector is a combined corner and edge detector which acts as a mature method that has been widely used in image processing. We can detect the corner and edge in the image using Harris detector. Similarly, we can get the shape and contour of objects and then we can process the features we detected. With the increasing amount of 3D data and the ability of capture devices to produce low-cost multimedia data, the capability to select relevant information has become an interesting research field. In 3D objects, the aim is to detect a few salient structures which can be used instead of the whole object, for applications like object retrieval, registration and mesh simplification. So, we can understand the 3D world through extraction and tracking of image features using a computer vision. Specially, how to detect the features in a video is a very critical problem. It is still unclear which features indicate useful interest points.

Many scholars have put forward a feature point extraction algorithm for video image. Wolf Kienzle approach the problem by learning a detector from examples (Kienzle *et al*., 2007). It records eye movements of human subjects watching video sequence and train a neural network to predict which locations are likely to movement targets. But the detector only outperforms current spatiotemporal interest point architectures on a standard classification dataset. Hedayati *et al*. (2013) reviews and compares the performance of five well-known detectors, SIFT, SURF, ORB, MSER and STAR, when combined in combination of with three common descriptors, SIFT, SURF and ORB. The results show that the SIFT and SURF detectors possess the most stable features for real-time video, with an overall accuracy of 80% under various conditions. But these methods have high time complexity for application. Hemati and Mirzakuchaki (2014) focus on recognizing the action of a person based on the appearance and motion information by constructing

spatio-temporal features in the "bag of keypoints" paradigm. It evaluates the performance of some local appearance detectors and descriptors along with the Histogram of Oriented Optical Flow as a motion descriptor. These features consist of SIFT, SURF and Harris-PHOG, each of which is concatenated with the HOOF of interest points. Willems *et al.* (2008) presents for the first time spatio-temporal interest points that are at the same time scale-invariant and densely cover the video content. It show that this can be achieved by using the determinant of the Hessian as the saliency measure and applying scale-space theory. Gauglitz *et al.* (2011) present a carefully designed dataset of video sequences of planar textures with ground truth which includes various geometric changes, such as lighting conditions and levels of motion blur and so on. It evaluate the impact of algorithm parameters, compare algorithms for both detection and description in isolation, as well as all detector-descriptor combinations as a tracking solution. Sipiran and Bustos (2011) propose an adaptive technique to determine the neighbourhood of a vertex, over which the Harris response on that vertex is calculated. The method is robust to several transformations which can be seen in the high repeatability values obtained using the SHREC feature detection and description benchmark. The experiments show that Harris 3D outperforms the results obtained by recent effective techniques such as Heat Kernel Signatures.

The problem we are addressing is that of using Harris to process the video. The video can be imagined as a composite of successive images (a video sequence), each image is one frame of video. Because the time interval between two frames is too short for the detection of human eyes, so, the video is continuous for our eyes. However, if we divided a video into frames and selected each frame, it is the same to an image, we do image processing to a single frame. Then, if we detect corners using Harris detector in a frame, we can get the corner information in a single frame. Moreover, if we deal with all the frames in a video, we can get corners or contour of objects in each frame and mark them out. Then we reorganized the frames into a video, so, that we can get the motion of this object. In this regard, we can do the target tracking which is very useful in the field of object recognition and objects change detection.

In this study, we expand the Harris corner detector from 2 dimensions to 3 dimensions. Firstly, we used Gaussian smoothing to eliminating gray change. Secondly, we used 3-D Harris detector to detect the corners in a video. Finally, we analysed the effect of the detector and made an assumption for its extended application.

THEORY OF 3-D HARRIS DETECTOR

Harris corner detector is a 2D corner detector operator which is based on local auto-correlation function, where it measures local changes in the image using the detection window to move by a pixel or sub-pixel in a different direction. It has a strong invariance of rotation, scaling, illumination changes and image noise.

Given a shift (4x,4y) and a point (x,y) in a 2D plane, the auto-correlation function is defined as Eq. 1:

$$c(x,y) = \Sigma_w \left[I(x_i, y_i) - I(x_i + \Delta x, y + \Delta y) \right]^2 \quad (1)$$

$$I(x_i + \Delta x, y_i + \Delta y) \approx I(x_i, y_i) - [I_x(x_i, y_i)] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (2)$$

where, $I(x_i,y_i)$ denotes the image function and (x_i,y_i) are the points in the detect window centered on (x_i,y_i), the shifted image is approximated by a Taylor expansion truncated to the first order terms.

where, $I_x(x_i,y_i)$, $I_y(x_i,y_i)$ are partial derivate at (x_i,y_i) in x and y direction, respectively:

$$c(x_i, y_i) = \sum_W [I((x_i, y_i) - I(x_i + \Delta x, y_i + \Delta y)]^2$$

$$= [\Delta x \, \Delta y] \begin{bmatrix} \sum_W I_x(x_i, y_i)^L & \sum_W I_x(x_i, y_i) \, I_y(x_i, y_i) \\ \sum_W I_x(x_i, y_i) \, I_y(x_i, y_i) & \sum_W I_y(x_i, y_i)^2 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}$$

$$[\Delta x \, \Delta y] C(x_i, y_i) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \quad (3)$$

where, matrix $C(x_i, y_i)$ is the structure tensor of the local neighborhood centered in (x_i, y_i) and sum means we should compute each pixel in the detect window, the sum them up.

We often use the gray value change to approximate $C(x_i, y_i)$, so in practical:

$$I_x = G(x+1,y)-G(x, y)$$

$$I_y = G(x, y+1)-G(x, y)$$

$$I_{xy} = G(x+1, y+1)+G(x,y)+2G(x+1,y) \quad (4)$$

where, $C(x_i, y_i)$ is a 2 by 2 matrix. Therefore, we can get two feature values of each tensor, $C(x_i, y_i)$ is an equation of ellipse, the smaller feature value represents the fastest luminance change direction which is the gradient direction. We have to do is to find the direction of gradient of each pixel and we get feature values of tensor at every pixel. Then, we select out the smaller one to form a new matrix according to its positions on the original image. Finally, we can obtain the corner strength image.

Then denote λ_1 and λ_2 as the feature value of $C(x_i, y_i)$. There are three cases to be considered:

- **Case 1:** If both λ_1 and λ_2 are small, so that the local auto-correlation function is flat
- **Case 2:** If one feature value is high and the other low, then only local shifts in one direction cause little change while in the orthogonal direction cause significant change, so, this would be the edge
- **Case 3:** If both feature values are high, then shifts in any direction will result in a significant increase, this indicates a corner

However, if we expand the 2D plane to a 3D video, we have an extra dimension called time t. Although, we think the video is a continuous image, it's time to connect each frame, so, we must deal with the time dimension and expand Harris detector into the video detector. Because of the extra dimension, we have to do is to add a dimension to the detector, in addition to the x and y direction of each pixel is detected, we should also detect the direction of t. This means that we should expand detectors from a window into a cube. At each pixel, we find that the variation of the gray values in the x, y and t. Therefore, the auto-correlation function becomes:

$$C(x,y,t) = \sum_W [I(x_i, y_i, t_i) - I(x_i + \Delta x, y_i + t_i + \Delta t)]^2$$

$$=[\Delta x\,\Delta y\,\Delta t]\begin{bmatrix} \sum_W I_x(x_i,y_i)^2 & \sum_W I_x(x_i,y_i)I_y(x_i,y_i) & \sum_W I_x(x_i,y_i)I_t(x_i,y_i) \\ \sum_W I_x(x_i,y_i)I_y(x_i,y_i) & \sum_W I_y(x_i,y_i)^2 & \sum_W I_y(x_i,y_i)I_t(x_i,y_i) \\ \sum_W I_x(x_i,y_i)I_t(x_i,y_i) & \sum_W I_y(x_i,y_i)I_t(x_i,y_i) & \sum_W I_t(x_i,y_i)^2 \end{bmatrix}\begin{bmatrix}\Delta x\\ \Delta y\\ \Delta t\end{bmatrix}$$

$$=[\Delta x\,\Delta y\,\Delta t]\,C(x_i,y_i,t_i)\begin{bmatrix}\Delta x\\ \Delta y\\ \Delta t\end{bmatrix} \quad (5)$$

After getting $C(x_i,y_i,t_i)$, we can calculate eigenvalues of tensor at each pixel, pick out the smallest from three and then form the corner strength cube.

However, since the Harris detector is based on the local auto-correlation function, we derive the derivative using this function at each point, so, we must ensure that the derivative is exist in each point. In the original image, some point in a certain direction may be a sudden gray value change, so there is not a derivative in this direction at this point. Before detection of the core algorithm, we must smooth image in order to eliminate a sudden the gray-scale change. In addition, we should also recognize that Harris detector is based on the local area to detect gray changes which means that it can detect changes in the value of all gray, including the point on an edge. However, we only need a corner point, so, we perform non-maximum suppression which can set a threshold value. When the angular intensity of the pixel is larger than this number, it can be classified as a corner point.

METHODOLOGY

Video process: In the three-dimensional model of the Harris detector, we use a cube-corner strength tensor to detect all of the pixels of the video signal which is also considered as a larger cube. However, in theory, its purpose is to make the theory easier to understand. In the procedure, we will divide video into frames to process, so, the first step is to divide the video into frames, each frame is a single image and use a structure variable to store these frames. This sample video is a 480×720×104, with RGB 3 bands. In order to make the program easier to run and debug, we select only 15 frames and a color band to test the algorithm. In a smooth process, the first two and the last two images don't smooth. The first two and the last three images is not treated by Harris detection loop. Because of these margin effects, we have only 6 images in the middle which is 5-10, can display the results.

Image smoothing: To guarantee that derivation exists in every point, smoothing is implemented. In this project we use Gaussian 1×5 filter window-Gauss = [0.10 0.3 0.45 0.3 0.10], it is convenient because all its elements' summation is 1, so, we don't need to divide the summation, this make the program efficient. In x and y direction, we can directly use conv2 function, but in t direction, because the pixels are not lie on the same frame. Therefore, we have to divide the Gauss window to let each element in Gauss window multiple the pixel in different frames, sum up as the gray value of the center pixel, then form a loop to process all the frames.

However, this is a phenomenon I met in smoothing. At the very beginning, because I believe the time between 2 frames is very short so that both the changes in motion and the value of gray, will be too small to perceive, so, I choose 15 frames with a interval of 5 to make the changes bigger. But after I get the result, I found though the changes are obvious between frames, the blur of the emotion is also obvious in each frame, as shown in Fig. 1.

Fig. 1(a-h): Original and smoothed image (a) 1st original, (b) 1st smoothed, (c) 2nd original, (d) 2nd smoothed, (e) 3rd original, (f) 3rd smoothed, (g) 4th original and (h) 4th smoothed

Figure 1 shows the scene of the animal in a river. we use Gaussian smoothing to process every frame image and generate smoothed image. After smoothing the image, it can eliminate the influence of some sharp information and improve the detection effect of edge and key points. It contributes to detect the detail information.

Drive corner strength: We imagine the detector as a 5×5 pixel cube in computing the corner strength, let the cube first go in x direction to detect the gray change of every pixel, then go in y direction to let the detector scan each pixel in a frame, then the last is t direction. We take 6 loops to finish this procedure. After the detector get a pixel's gray changes in all direction, form the M matrix and select the smallest eigenvalue, then fill this eigenvalue at the corresponding position in a new cube. After all the video pixel are processed, we get the corner strength cube. In the program I define this cube as a 3 dimension matrix, x and y represent the pixel coordinate in a frame while t represents each frame.

Non-maximum suppression: There are two main tasks in non-maximum suppression. The first task is to use the algorithm to detect the real corner according to the threshold we set. We set the threshold and detector size and then we use the core part of non maximum suppression, the ordfilt3 3D filter. The filter will set the pixels in the detectable space as 1 which we manually set the pixel in the undetectable space as 0 and then the filter will scan each pixel to detect its corner strength, if the corner strength at this pixel is bigger than threshold, meanwhile it is the biggest one in the detector. The filter will pick out that pixel as a 'real corner' and remain the value of that pixel as 1 while if the strength corner is smaller than the threshold, or its corner strength is not the biggest one in the detector, it won't be considered as a true corner and the value will be set as 0. The 1s and 0s is not the gray value of that pixel, just for our convenient to find out that pixel. The second task is to record the coordinate of these real corner for later use.

RESULTS AND DISCUSSION

Figure 2 is the detecting results of a few frames image. It have different results under the same scenario. This may be the reason for the image of the light changes. The strong contrast of the image have feature points to some more and less weak contrast of image have a few feature points. At the same time, it is possible that the noise caused the result changes by processing each frame. From the result strength corner image, we can see that the contour of bear, birds, rocks are clear, it may be caused by two reasons as follow.

On the one hand, Harris detector is based on the detection of the gray value changes which select the smallest feature value of a pixel as its corner strength. The corner has a gray value change in every direction while the edge also has a change in one direction, I think perhaps may be there is a noise affection, or the edge is very coarsely, that is to say the edge is like many points, so after the detection, the feature value is small enough and the corner strength is big enough to be shown on the corner strength image. Therefore, the edge is also detected.

On the other hand, the reason has something to do with the sample video. In this video, there is not much corners, most of the features are edge. The only feature that we can classify as the corner is the water in the river, where the tide's shape is irregular so, that the detector can make it as the corner. The results show that the detector detected the corner of the tide.

Table 1 and Fig. 3 show, respectively the number and distribution of feature points. We can find some different information from Table 1. Some image did not detect the feature points and some image have more and similar number feature points. The total of characteristic points is 507 which the results can express the change of the target in the scene. From Fig. 3, some information can be found that the distribution of feature points

Fig. 2(a-d): Results of corner detector (a) 1st, (b) 2nd, (c) 3rd and (d) 4th frame, respectively

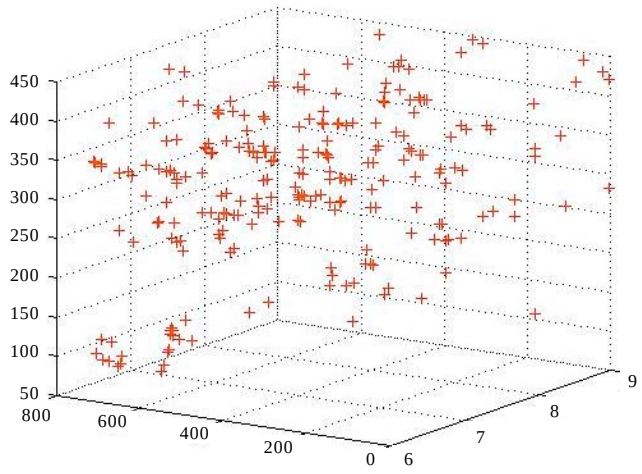

Fig. 3: Corners in the video space

Table 1: Statistics of the results

Image No.	Corner No.
1st	0
2nd	173
3rd	188
4th	146
Total	507

is relatively concentrated that imply a bear. It represents the key target in the scene. A outlier area shows that some of the other targets. It also have some non target feature points.

In order to compare our method with the state of the art, we selected three recent methods for detecting interest points on video: 3D-HOG, 3D-SURF and 3D-SIFT. In the following, we specify the procession used for these methods:

- The 3D-HOG generalizes the HOG concepts to 3D through viewing videos as spatio-temporal volumes in order to recognizing action (Klaser *et al*., 2008; Dang *et al*., 2011). It have extended integral images to integral videos for 3D gradient computation and presented a quantization method for 3D orientations. It includes gradient computation in arbitrary spatial and temporal scales, orientation quantization of 3D gradients with projection, histogram and descriptor computation. But it detects interesting points with much higher variability for blurry video
- The 3D-SURF detects spatio-temporal interest points with the determinant of the Hessian at some positions and some scales (Willems *et al*., 2008; Knopp *et al*., 2010). It builds on integral videos in order to make the problem of position and scale tractable which converts a video containing F frames of dimension (W, H) into an integral video structure where an entry at location (x,y,t) holds the sum of all pixels in the rectangular region. It uses the determinant of the Hessian in order to achieve the search of scale-space. To describe the interest points, it makes use of an extend version of the SURF descriptor with spatial scale and temporal scale. Nevertheless, there are a big problem of the accuracy of positioning resulted from the gradient of local area pixel

- The 3D-SIFT computes the overall orientation of the neighborhood (Scovanner *et al*., 2007; Abdul-Jauwad *et al*., 2012). Then, it create the sub-histograms which will encode 3D SIFT descriptor. For a given cuboid, spatio-temporal gradients are computed for each pixel. All pixels vote into a three dimension grid of histograms of oriented gradients. For orientation quantization, gradients are represented in polar coordinates that are divided into a (8, 4) histogram bins. This leads to some problems: Singularities and inefficiency

Harris corner detection algorithm is a corner feature extraction operator. It is based on image grey signal, proposed by Harris and Stephens (Harris and Stephens, 1988). In this study, the Harris corner detection operator is extended from 2D-3D corner detection in video. The 3D Harris obtained the best results with respect to above several questions for our video. Firstly, there is a total predominance of our method with regard to edge information. Secondly, we outperforms the rest in positioning precision for local area pixel. Finally, it is interesting that our method is also the best in efficiency for video.

CONCLUSION

In this study, we introduce a method of 3-D Harris detector to detect feature points for a video. It identifies the corner and edge from every frames. This method is suitable for the dynamic changes of the image. It contributes to find the change of scene which can apply to the change detection of 3d video.

ACKNOWLEDGMENTS

Thanks to Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation (No. 20111211N, Z201019), National Key Technology Support Program (No. 2012BAK13B06), Beijing Municipal Education Commission General Program (No. 03058314105) and Beijing Natural Science Fund Project (8142014).

REFERENCES

Abdul-Jauwad, S.H., R. Ullah and F. Ullah, 2012. A simple method to recover 3-D rigid structure from motion using SIFT, RANSAC and the Tomasi-Kanade factorization. Int. J. Comput. Sci. Issues, 9: 122-128.

Dang, L., B. Bui, P.D. Vo, T.N. Tran and B.H. Le, 2011. Improved HOG descriptors. Proceedings of the 3rd International Conference on Knowledge and Systems Engineering, October 14-17, 2011, Hanoi, pp: 186-189.

Gauglitz, S., T. Hollerer and M. Turk, 2011. Evaluation of interest point detectors and feature descriptors for visual tracking. Int. J. Comput. Vision, 94: 335-360.

Harris, C. and M. Stephens, 1988. A combined corner and edge detector. Proceedings of the 4th Alvey Vision Conference, Volume 15, August 31-September 2, 1988, Manchester, UK., pp: 147-151.

Hedayati, M., W.M. Diyana, W. Zaki, A. Hussain and M.A. Zulkifley, 2013. Performances of invariant feature detectors in real-time video applications. Proceedings of the 3rd International Visual Informatics Conference on Advances in Visual Informatics, November 13-15, 2013, Selangor, Malaysia, pp: 193-205.

Hemati, R. and S. Mirzakuchaki, 2014. Using local-based Harris-PHOG features in a combination framework for human action recognition. Arabian J. Sci. Eng., 39: 903-912.

Kienzle, W., B. Scholkopf, F.A. Wichmann and M.O. Franz, 2007. How to find interesting locations in video: A spatiotemporal interest point detector learned from human eye movements. Proceedings of the 29th DAGM Symposium on Pattern Recognition, September 12-14, 2007, Heidelberg, Germany, pp: 405-414.

Klaser, A., M. Marsza³ek and C. Schmid, 2008. A spatio-temporal descriptor based on 3d-gradients. Proceedings of the 19th British Machine Vision Conference, September 1-4, 2008, Leeds, pp: 995-1004.

Knopp, J., M. Prasad, G. Willems, R. Timofte and L. Van Gool, 2010. Hough transform and 3D SURF for robust three dimensional classification. Proceedings of the 11th European Conference on Computer Vision, September 5-11, 2010, Heraklion, Crete, Greece, pp: 589-602. September 5-11, 2010, Heraklion, Crete, Greece, pp: 589-602.

Scovanner, P., S. Ali and M. Shah, 2007. A 3-dimensional sift descriptor and its application to action recognition. Proceeedings of the 15th International Conference on Multimedia, September 23-28, 2007, Bavaria, Germany, pp: 357-360.

Sipiran, I. and B. Bustos, 2011. Harris 3^D: A robust extension of the Harris operator for interest point detection on 3^D meshes. Visual Comput., 27: 963-976.

Willems, G., T. Tuytelaars and L. van Gool, 2008. An efficient dense and scale-invariant spatio-temporal interest point detector. Proceedings of the 10th European Conference on Computer Vision, October 12-18, 2008, Marseille, France, pp: 650-663.