Multimed Tools Appl
DOI 10.1007/s11042-015-2694-2

Tracking feature extraction techniques with improved SIFT for video identification

Ruichen Jin[1] · **Jongweon Kim**[2]

Received: 12 February 2015 / Revised: 25 April 2015 / Accepted: 12 May 2015

Abstract This paper presents a method for tracking of object movements and detecting of feature to identify video content using improved Scale-Invariant Feature Transform (SIFT). SIFT can robustly identify objects even among clutter and under partial occlusion, because the SIFT feature descriptor is invariant to uniform scaling, orientation, and also partially invariant to affine distortion and illumination changes. Even if the video drops frames or attacked, our method can extract the features. In our method we detect the video features from tracking the object's movement and make a dataset with feature sequences to identify video. In contrast to the existing tracking techniques, our method recognized reliable object coordinate. The developed algorithm will be an essential part of a completely tracking and identification system. To evaluate the performance of the proposed approach, we was experimenting with several genres of video. Compare with the original SIFT algorithm, we reducing up to 5 % in processing time was achieved for matching. Also appoint the position of the object area in tracking method make the proposed method automatic, fast and effective.

Keywords Feature detection · Tracking · Video feature · SIFT · Identification · Torrent

1 Introduction

Recently, due to the widespread use of PC as multimedia systems, the illegal distribution of digital content has become a social issue [5, 7–9]. Especially, a lot of videos are distributed without the permission of copyrighter. The formal example is distributing by Torrent protocol. Multimedia matching technology is widely used in the field of image processing such as image

✉ Jongweon Kim
jwkim@smu.ac.kr

Ruichen Jin
kimyejin0602@gmail.com

[1] Department of Copyright Protection, Sangmyung University, Seoul, Korea

[2] Department of Contents and Copyright, Sangmyung University, Seoul, Korea

Published online: 22 May 2015

mosaic and fusion, target recognition and tracking, photogrammetry, remote sensing image retrieval, and machine vision. The current methods for image matching is divided into pixel based matching and feature based matching. In recent years, image local invariant feature extraction method has got very fast development.

In massive database, matching the video feature and claiming own copyright need the technologies with speed and accuracy. In this paper, we propose an improved method for tracking of the object movements and detecting the feature to identify video. Compare with original SIFT, the proposed algorithm greatly reduce the dimension of feature with the simple calculation. We tracking the object with P-N learning. The position of object area is appointed by the improved SIFT.

The rest of this paper is organized as follows: Sesion2 summarizes the existing matching and feature detection methods. Session 3 presents an overview of our proposed method. Session 4 evaluates the performance of our method through a number of experiments. Session5 briefly concludes this paper.

2 Related work

2.1 Methods for optical flow

The simple way to track an object is finding the object area in the next frame. Horn and Schunk [4] assumed a constant that does not change depending on the brightness value of a pixel in the gray level patterns in time and added the constraint that velocity distribution is changed smoothly from all pixels of the image. In other words, they assumed the movement of pixels is not great because the time of the interval between frames is very short. The disadvantage of this method is slow, because it performs the optical flow calculated from all pixels.

Currently, the most widely used methods of optical flow are proposed by Lucas and Kanade [12]. To obtain the vector to minimize the square error of the optical-flow equation, calculate a vector for each pixel by local least squares within a specific range of the neighbor pixels in the window.

2.2 Methods for tracking

The methods for tracking analyze the video frame and perform a progressive video tracking algorithm to perform the movement of the target between the frames. The algorithm has a variety of strengths and weaknesses. It is important to consider when choosing an algorithm for the intended use. As well as the target localization, expression and the data association filter has two main components of the visual tracking system.

2.3 Methods for SIFT

The SIFT algorithm (Scale Invariant Feature Transform) proposed by Lowe [10] is an approach for extracting distinctive invariant features from images. It has been successfully applied to a variety of computer vision problems based on feature matching including object recognition, pose estimation, image retrieval and many others. However, in real-world applications there is still a need for improvement of the algorithm's robustness with respect to the

correct matching of SIFT features. After a series of studies, he tried to put the invariance effective for all affine transformations. Research in paper [11] shows SIFT features may be applied to a range of affine transformation. After scale space extrema are detected (their location being shown in the uppermost image) the SIFT algorithm discards low contrast keypoints (remaining points are shown in the middle image) and then filters out those located on edges. Resulting set of keypoints is shown on last image.

Simply summarized the operation of the SIFT as follows:

1. Scale-space extrema detection—Find locally bouncing areas—parts of characteristic, which is represented by the features of the object.
2. Keypoint Location—The process of selecting the features section to find out the previously characterized. It is determined based on values of pixel (Intensity), position or size. It is prefer to extract corners rather than edges.
3. Orientation Assignment—In this step, each keypoint is assigned one or more orientations based on local image gradient directions. This is the key step in achieving invariance to rotation as the keypoint descriptor can be represented relative to this orientation and therefore achieve invariance to image rotation.
4. Keypoint Descriptor—This ensured invariance to image location, scale and rotation. This step is performed on the image closest in scale to the keypoint's scale (Fig. 1). After scale space extrema are detected (their location being shown in the uppermost image) the SIFT algorithm discards low contrast keypoints (remaining points are shown in the middle image) and then filters out those located on edges. Resulting set of keypoints is shown on last image.

The method of principal component analysis histogram reduce the dimensionality of feature. H. Bay proposed SURF algorithm can speed up robust features, algorithms in [1, 6]. A lot of other feature types have also been used for image recognition, image contour, such as phase, color etc. Local points combined with these features can enhance the robustness of recognition, image recognition system in the future is likely to adopt the multi feature combination.

3 Material and methods

SIFT flow is a method to measure the similarity between scene images by optimizing SIFT feature dense flow [2]. SIFT feature is invariant for rotations, scale changes, and illumination

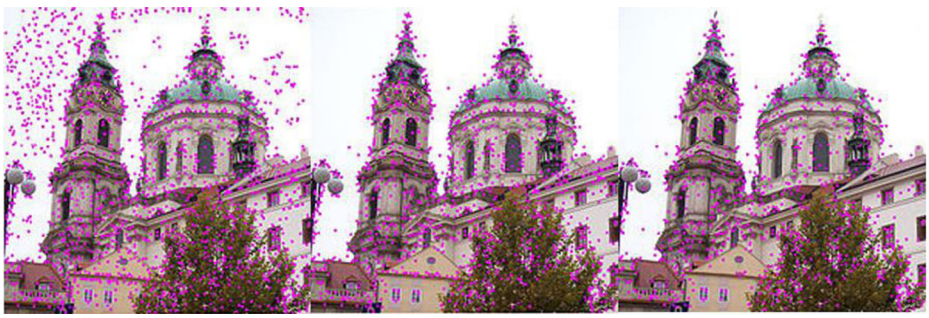


Fig. 1 Visualization of SIFT image

changes and it is often used for object recognition. The scale-invariant features are efficiently identified by using a staged filtering approach. In this approach the first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame.

Keypoints are detected D(x, y, σ) which is the difference of smoothed images L(x, y, σ). L(x, y, σ) is obtained from the convolution of variable scale Gaussian with the input image I(x, y).

$$\mathrm{L}(\mathrm{x},\mathrm{y},\sigma) = \mathrm{G}(\mathrm{x},\mathrm{y},\sigma)*\mathrm{I}(\mathrm{x},\mathrm{y})$$

It is performed difference between scales σ, and the number of images are obtained. Local extrema are detected from these images by comparing neighborhood of pixels within a set of images, and if it is an extremum, the pixel is detected as the keypoint.

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} exp\left(-\left(x^2+y^2\right)\Big/2\sigma^2\right)$$

$$D(x,y,\sigma) = L(x,y,k\sigma) - L(x,y,\sigma)$$

The gradient magnitude m(x, y) and orientation θ(x, y) at each pixel of image that the keypoints are detected using as the following expressions:

$$\mathrm{m}(\mathrm{x},\mathrm{y}) = \left\{[\mathrm{L}(\mathrm{x}+1,\mathrm{y})-\mathrm{L}(\mathrm{x}-1,\mathrm{y})]^2 + [\mathrm{L}(\mathrm{x},\mathrm{y}+1)-\mathrm{L}(\mathrm{x},\mathrm{y}-1)]^2\right\}^{1/2}$$

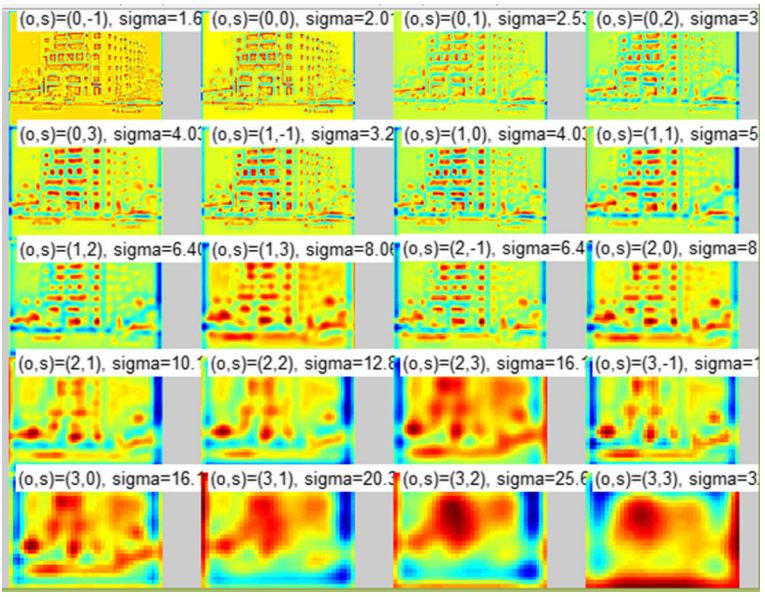


Fig. 2 Visualization of SIFT image

Fig. 3 Keypoints selection

$$\theta(x,y) = tan^{-1}\left[\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right]$$

A minimum Euclidean distance vector and the second smallest Euclidean distance ratio is less than 0.8 on the basis. Both cosine similarity is further determined whether the vector is greater than a certain threshold value experience, if greater then start matching. Experience threshold k can be obtained by experiment. Cosine similarity is the cosine of the angle between

Fig. 4 Feature matching

Table 1 Information of feature detection

No.	Name	Total keypoints	Matching points	t/s
1	Snowman	267	38	0.184
2	Man	331	12	0.153
3	Skell	128	22	0.093
4	Building	69	11	0.164
5	Book	118	21	0.230
6	Onepiese	519	69	0.148
7	Smog	220	36	0.253
8	Bike	173	66	0.156

two vectors, which is not subject to the reference coordinate system rotation, zoom effects, and the standardization of the length of the vector. For two vectors x and y:

$$S(x,y)-\frac{(x,y)}{||x||\left|\left|y\right|\right|}-\frac{\sum_{i=1}^{n}x_i y_i}{\left(\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2\right)^{1/2}}$$

If the locations of potential keypoints are found, they have to be refined to get more accurate results. We used Taylor series expansion of scale space to get more accurate location of extrema.

Identify keypoints between the two frames are matched to the nearest neighbors. However, in some cases, a close match can be very close to the work. It can cause noise or other reasons. In this case, the ratio of the second closest to the close distance is taken as. If it is greater than 0.8, it is rejected. It eliminated around 90 % of false matches while discards only 5 % correct matches.

Figure 2 shows the visualization of SIFT for image. Figure 3 shows the keypoints selection.

Detect the feature of video from tracking the object's movement and make a dataset with feature sequence for video identification. We propose to use the P-N learning as a tracking measure [13]. It is also efficient to find the feature points, and even faster [3, 14].

P-N learning consists of four parts: (1) to be learning a classifier; (2) the training sample set—some known class label samples; (3) supervised learning—an intensive training sample

Fig. 5 Detect keypoints and identify the object

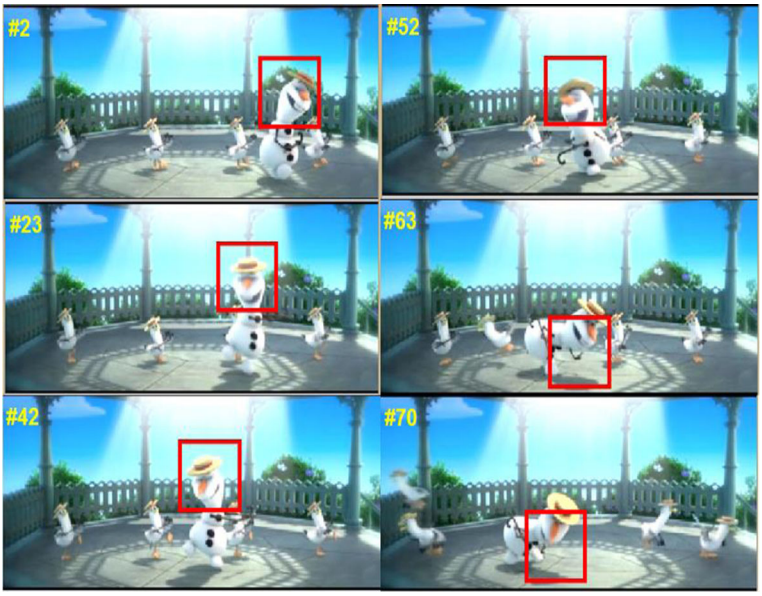


Fig. 6 Pointing the coordinate (manual) and tracking the object

from the training of the classifier methods; (4) P-N experts- in the learning process for generating positive (training) expression function and negative samples (training) sample.

Most paper pointing the coordinate and setting the area to tracking the object. In order to reduce the time, we use the improved SIFT to catch the object. Figure 3 shows the detected keypoints. We catch the area with occupying a large portion of keypoints (Fig. 4).

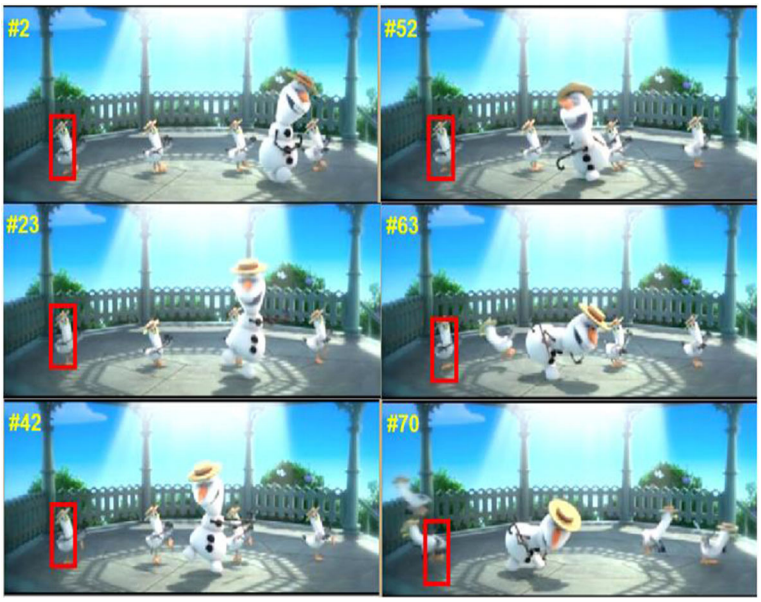


Fig. 7 Object recognition (automatic) and tracking the object

Table 2 Comparison of false matching rates obtained from Using improved algorithm and original algorithm

Video feature dataset	Percentage(%)	
	Orignal	Improved
1	6.1	2.6
2	5.3	3.6
3	17.5	9.3
4	21.4	7.4

Table 1 shows matching result including the information of keypoints detection. There are enough matching points to recognize object. The Snowman total has 267 keypoints. Matching with the close frame, they have 38 matching points, and it costs 0.184 s. The cost time is between about 0.1~0.3 s.

4 Experimental results

To evaluate the performance of the proposed method, we have construct a features dataset from several genres video clip for these experiments. Figure 5 shows the match points between the adjacent frames and specify the area which the area is clustering with matching point.

Figure 6 shows the traditional P-N learning tracking method, we pointed the coordinate and the size of rectangle manually. Figure 7 also use the P-N learning tracking method, but the object is detected using proposed method.

Those figures shown the different extracted object, even if there are many cases that the objects to extract are the same. To recognize the object by algorithm can make matching rate higher. In addition, although not conspicuous things can be extracted with an object.

Test by using classical algorithms, adding the cosine similarity constraint, direction, ultimately shows improving of consistency constraint matching algorithm. Tables 2 and 3 show the video library 4 groups' results. Table 2 shows the comparison of false matching rates obtained from Using improved algorithm and original algorithm. Table 3 shows the comparison of matching time using improved algorithm and original algorithm. The size of video sample is 800×640 pixels. The Fig. 8 shows the chart of result. The proposed method have improvement in false matching rates and the matching time is reduced as well.

Table 3 Comparison of matching time using improved algorithm and original algorithm

Video feature dataset	t/s	
	Orignal	Improved
1	9.821	10.514
2	5.341	4.501
3	5.621	4.368
4	5.551	5.271

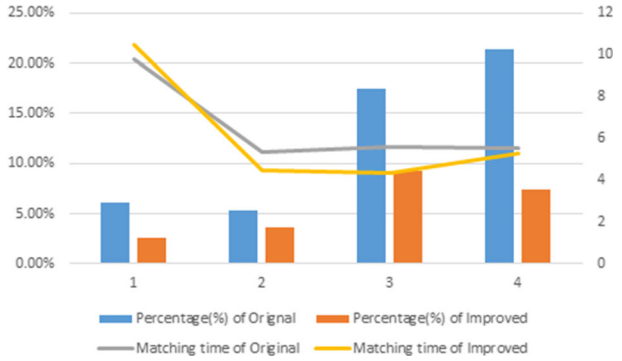


Fig. 8 Chart of results

5 Conclusion

In conclusion, we have introduce a new effective method for detecting feature for video identification. Unlike the classical SIFT, the proposed algorithm greatly reduce the dimension of feature with the simple calculation.

With respect to the original algorithm, the improved algorithm have better results for rotation and scaling, noise, blur, light and other transformation and the average mismatching rate reduced by 10 to 20 %. For perspective transformation, the improved algorithm cannot show obvious improving for small-scale perspective transformation (40 ° or less). Mismatching rate reduced by an average of 5 % or less. When the viewing angle exceeds 40 °, the original algorithm and improved algorithm mismatch rate of over 70 %.

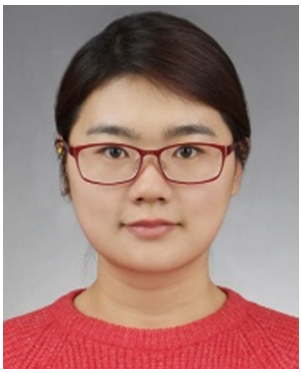
Experimental results show that this method can be the basis of ensuring the correct match time and the match points, image rotation and scaling, noise, blur, lighting changes and small-scale perspective transformation model is robust and effectively reducing the false match rate. But for perspective transform how large high-precision matching, further research is needed.

Acknowledgments This research project was supported by the Ministry of Culture, Sports and Tourism (MCST) and the Korea Copyright Commission in 2014.

References

1. Bay H, Tuytelaars T, Van Gool L (2006) SURF: speeded up robust features, Proceedings of the ninth European Conference on Computer Vision, May 2006
2. Ce L, Jenny Y, Antonio T (2011) SIFT flow: dense correspondence across scenes and its applications, IEEE Transactions on Pattern Analysis and “Machine Intelligence”, 33(5)
3. Chapelle O, Schokopf B, Zien A (2006) Semi-surpervised learning. MIT Press, Cambridge
4. Horn BKP, Schunk BG (1981) Determining optical flow. Artif Intell 17:185–203
5. Jin R, Kim J (2012) A digital watermarking scheme using hologram quantization, SIP2012 342:39–46
6. Ke Y, Sukthankar R (2004) PCA-SIFT: a more distinctive representation for local image descriptors, Computer Vision and Pattern Recognition
7. Kim J, Kim N, Lee D, Park S, Lee S (2010) Watermarking two dimensional data object identifier for authenticated distribution of digital multimedia contents. Signal Process Image Commun 25:559–576
8. Lee Y, Kim J (2011) Robust blind watermarking scheme for digital images based on discrete fractional random transform. Commun Comput Inf Sci 263(139145)

9. Li D, Kim J (2012) Secure image forensic marking algorithm using 2D barcode and off-axis Hologram in DWT-DFRNT domain. Appl Math Inf Sci (AMIS) 6(2S):513–520
10. Lowe DG (1999) Object Recognition from local scale-invariant features. International Conference on Computer Vision, Corfu, Greece (Sep. 1999), 1150–1157
11. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
12. Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision, In Proceedings of the International Joint Conference on Artificial Intelligence, 674–679
13. Zdennek K, Krystian M, Jiri M. Tracking-learning detection
14. Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3: 1–130



Ruichen Jin received her B.S. degree in Computer Science and Technology from YanBian University, China, in 2011. She is currently pursuing the Ph.D. degree in Copyright Protection, Sangmyung University, Korea. Her research interests are digital watermarking, multimedia forensics, digital signal processing, and information security.

Jongweon Kim received the Ph.D. degree from University of Seoul, major in signal processing in 1995. He is currently a professor of Dept. of Contents and Copyright at Sangmyung University in Korea. He has a lot of practical experiences in the digital signal processing and copyright protection technology in the institutional, the industrial, and academic environments. His research interests are in the areas of copyright protection technology, digital rights management, digital watermarking, and digital forensic marking.