INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

SIFT FOR LARGE SCALE IMAGE SEARCH

B. Mathan kumar[1], N. Suresh kumar[2], T. Chitra kumar[3]

[1]Assistant Professor, Department of MCA, PSNA College of Engineering &Technology, Dindigul.
mathan81@gmail.com
[2]Assistant Professor (Sl.Gr), Department of IT, Sri Ramakrishna Engineering College, Coimbatore.
nsuresh2@gmail.com
[3]Assistant Professor, Department of IT, Sri Ramakrishna Engineering College, Coimbatore.
chithrakumar.thangaraj@srec.ac.in

Abstract: - Image search in image compression includes the mapping of visual features into compact binary codes. It needs more memory space for storage. So in order to reduce the memory cost for storage, there is a need for encoding high-dimensional data as compact binary codes. The similarity is also an important consideration in large scale image search. The computational efficiency can be computed based on the computation of similarity. It is measured by using hamming distance. Signals are used to detect the feature. It is achieved by taking correlation between two signals (cross correlation).For high frequency Domain expression the normalized form of correlation (correlation coefficient) preferred in template matching does not have a correspondingly simple and efficient frequency domain expression. To avoid this normalized cross-correlation has been computed in the spatial domain. As the computational cost of spatial domain convolution is high, more inexact but fast spatial domain matching methods have also been developed. To obtain normalized cross correlation from transform domain convolution the Fast Normalized Cross Correlation Algorithm is introduced. This new algorithm provides an order of magnitude speedup over spatial domain computation of normalized cross correlation.

Keywords: Image, Cross correlation, spatial domain, coefficient, patterns.

1. INTRODUCTION

In recent years, content-based image search has attracted more and more attentions in the multimedia and the computer vision community [1]–[17]. Many approaches are based on the Bag-of-Visual-Words (BoVW) model [8] and adopt the invariant local features for image representation. In the BoVW model, an image is represented by a visual word vector. The visual words are usually generated by clustering the extracted local features. Some widely used unsupervised clustering algorithms are standard k-means, hierarchical k-means (HKM) [7].

2. THE CORRELATION BETWEEN TWO SIGNALS (CROSS CORRELATION)

Is a standard approach to feature detection [6,7] as well as a component of more sophisticated techniques (e.g. [3]). Textbook presentations of correlation describe the convolution theorem and the attendant possibility of

efficiently computing correlation in the frequency domain using the fast Fourier transform. Unfortunately the normalized form of correlation (correlation coefficient) preferred in template matching does not have a correspondingly simple and efficient frequency domain expression. For this reason normalized cross-correlation has been computed in the spatial domain (e.g., [7], p. 585). Due to the computational cost of spatial domain convolution, several inexact but fast spatial domain matching methods have also been developed [2]. This paper describes a recently introduced algorithm [10] for obtaining normalized cross correlation from transform domain convolution. The new algorithm in some cases provides an order of magnitude speedup over spatial domain computation of normalized cross correlation. Since we are presenting a version of a familiar and widely used algorithm no attempt will be made to survey the literature on selection of features, whitening, fast convolution techniques, extensions, alternate techniques, or applications. The literature on these topics can be approached through introductory texts and handbooks and recent papers such as [1]. Nevertheless, due to the variety of feature tracking schemes that have been advocated it may be necessary to establish that normalized cross-correlation remains a viable choice for some if not all applications.

The rest of the paper is organized as follows. Section 2 reviews the pipeline of content-based large-scale image search system with local features. Section 3 discusses the proposed Fast Normalized Cross Correlation Algorithm. Our transforming domain computation is introduced in Section 4 Phase correlation and normalization are given in Section 5 and Section 6. Finally, the conclusion is made in Section 7. In content-based large-scale image retrieval with local features, the Bag-of-Visual-Words (BoVW) model has been widely adopted. Generally, most approaches follow a pipeline which consists of several key steps, including feature representation, feature quantization, image indexing, image scoring and post-processing. In this section, we make a review of the pipeline and discuss related works in each step. Invariant local features have been popularly adopted for image representation owing to its invariance to various transformations and robustness to occlusions and background changes. The extraction of local features usually involves two steps, namely interest point detection and feature description. The interest point detection identifies some key points that have high repeatability over various changes. The commonly used interest point detectors include Difference of Gaussian (DOG)[5].

Then, a descriptor is constructed to capture the visual appearance of the local region corresponding to a interest point. The descriptor is usually designed to be invariant to rotation and scale changes and robust to affine distortion and the addition of noise, etc. Some commonly used descriptors include FNCC, SURF [5], and some recently developed descriptors.

To obtain a compact representation of an image for scalable indexing and retrieval, after extracting the local features, a visual vocabulary is usually built and the local features are quantized to visual words. The visual vocabulary is generated by unsupervised clustering algorithms, such as hierarchical k-means (HKM) [7], approximate k-means (AKM). With the visual vocabulary defined, each local feature is quantized to a visual word in the visual vocabulary. Usually, to speed up quantization process, approximate nearest neighbor (ANN) approaches are adopted, such as k-d tree, vocabulary tree [7]. A scalar quantization approach is proposed to suppress the quantization error. After the features are quantized to visual words, an image can be represented by a visual word vector [8].

The similarity between two images can be measured by the L1 or L2 normal distance between their visual word vectors. And inspired by the success of text search engines, the inverted file structure has been successfully used for large-scale image search [8],[10],. In the inverted file structure, each visual word is followed by a list of entries. Each entry records the image ID and some other clues to verify the feature matching, such as geometric clues and binary signatures.

As the BoVW model ignores the spatial context information between local features, some researchers propose to conduct geometric verification to the ranked candidates list returned by the BoVW model. A transformation model between the query image and the candidate image is estimated and those matches that do not fit the model well are filtered out.

Figure (a) Image matching in image search

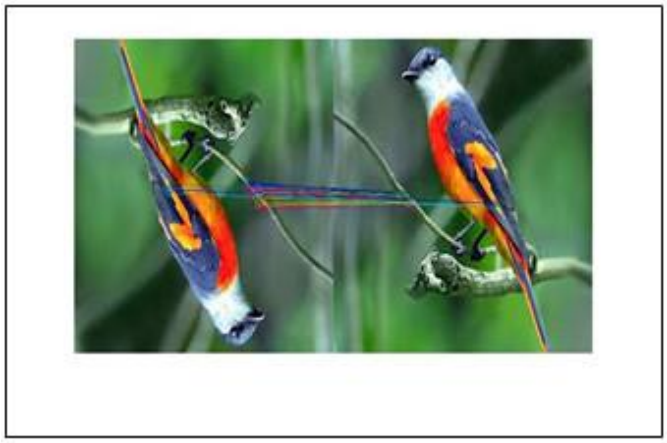

Figure (b) Cross Indexing of matched images

Unlike the above geometric verification approaches, the bag-of-spatial features [10] explicitly embeds the spatial context into the representation based on visual words. Some works explore the semantic information of images. IA sparse graph-based semi-supervised learning approach is used to inferring images' semantic concepts from community-contributed images and their associated tags. The semantic gap measure is introduced into the active learning process to handle the user interaction.

3 FAST NORMALIZED CROSS CORRELATION ALGORITHM

Template Matching by Cross-Correlation: The use of cross-correlation for template matching is motivated by the distance measure (squared Euclidean distance)

$$d^2 f, t\,(u,v)=\sum [f\,(x,y)-t(x-u,y-v)]^2$$

(where f is the image and the sum is over x,y under the window containing the feature t is positioned at u,v). In the expansion of d^2

$$d^2 f, t\,(u,v) = \sum\ [f^2(x,y)-2\ \ f\,(x-y)\,t\,(x-u,y-v) + t^2(x-u,y-v)].....(1)$$

the term $\sum t^2(x-u,y-v)$ is constant. If the term $\sum f^2(x,y)$ is approximately constant then the remaining cross-correlation term $c(u,v)= \sum f(x,y)t(x-u,y-v)$ is a measure of the similarity between the image and the feature.

There are several disadvantages to using (1) for template matching: If the image energy $\sum f^2\,(x, y)$ varies with position, matching using (1) can fail. For example, the correlation between the feature and an exactly matching region in the image may be less than the correlation between the feature and a bright spot. The range of c(u,v) is dependent on the size of the feature.

Equation (1) is not invariant to changes in image amplitude such as those caused by changing lighting conditions across the image sequence. The correlation coefficient overcomes these difficulties by normalizing the image and feature vectors to unit length, yielding a cosine-like correlation coefficient

$$(u,v) = \frac{\sum_{x,y} [f(x,y)-f_{u,v}][t(x-u,y-v)-t]}{\{\sum_{x,y}[f(x,y)-f_{u,v}]^2 \sum_{x,y}[t(x-u,y-v)-t]^2\}^{0.5}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where t is the mean of the feature and $f_{u,v}$ is the mean of f (x,y) in the region under the feature. We refer to (2) as normalized cross-correlation.

4 FEATURE TRACKING APPROACHES AND ISSUES

The feature tracking approach uses the normalized cross-correlation (NCC) and is not the ideal approach to feature tracking because it is not invariant with respect to imaging scale, rotation, and perspective distortions. The problems have been addressed in various schemes which represents the feature tracking techniques which includes the NCC as a component. This paper provides the choice of matching the features with the help of NCC with various alternate approaches. For this reason NCC is recommended for feature tracking since it point out some of the issues that is involved in feature tracking, and it represents that NCC is a efficient means for large scale image search.

SSDA. The basis of the sequential similarity detection algorithm (SSDA) [2] is the observation that full precision is only needed near the maximum of the cross-correlation function, while reduced precision can be used elsewhere. The authors of [2] describe several ways of implementing reduced precision. An SSDA implementation of cross-correlation proceeds by computing the summation in (1) in random order and uses the partial computation as a Monte Carlo estimate of whether the particular match location will be near a maximum of the correlation surface. The computation at a particular location is terminated before completing the sum if the estimate suggests that the location corresponds to a poor match. The SSDA algorithm is simple and provides a significant speedup over spatial domain cross-correlation. It has the disadvantage that it does not guarantee finding the maximum of the correlation surface. SSDA performs well when the correlation surface has shallow slopes and broad maxima. While this condition is probably satisfied in many applications, it is evident that images containing arrays of objects (pebbles, bricks, other textures) can generate multiple narrow extreme in the correlation surface and thus mislead an SSDA approach. A secondary disadvantage of SSDA is that it has parameters that need to determined (the number of terms used to form an estimate of the correlation coefficient, and the early termination threshold on this estimate).

Gradient Descent Search: If it is assumed that feature translation between adjacent frames is small then the translation can be obtained by gradient descent [12]. Successful gradient descent search requires that the interframe translation be less than the radius of the basin surrounding the minimum of the matching error surface. This condition may be satisfied in many applications. Images sequences from hand-held cameras can violate this requirement, however: small rotations of the camera can cause large object translations. Small or (as with SSDA) textured templates result in matching error surfaces with narrow extrema and thus constrain the range of interframe translation that can be successfully tracked. Another drawback of gradient descent techniques is that the search is inherently serial, whereas NCC permits parallel implementation.

Snakes: Snakes (active contour models) have the disadvantage that they cannot track objects that do not have a definable contour. Some ``objects'' do not have a clearly defined boundary (whether due to intrinsic fuzzyness or due to lighting conditions), but nevertheless have a characteristic distribution of color that may be trackable via cross-correlation. Active contour models address a more general problem than that of simple template matching in that they provide a representation of the deformed contour over time. Cross-correlation can track objects that deform over time, but with obvious and significant qualifications that will not be discussed here. Cross-correlation can also easily track a feature that moves by a significant fraction of its own size across frames, whereas this amount of translation could put a snake outside of its basin of convergence.

Wavelets and other multi-resolution schemes: Although the existence of a useful convolution theorem for wavelets is still a matter of discussion (e.g., [11]; in some schemes wavelet convolution is in fact implemented using the Fourier convolution theorem), efficient feature tracking can be implemented with wavelets and other multi-resolution representations using a coarse-to-fine multi-resolution search. Multi-resolution techniques require, however, that the images contain sufficient low frequency information to guide the initial stages of the search.

Each of the approaches discussed above has been advocated by various authors, but there are fewer comparisons between approaches. Reference derives an optimal feature tracking scheme within the gradient search framework, but the limitations of this framework are not addressed. An empirical study of five template matching algorithms in the presence of various image distortions [4] found that NCC provides the best performance in all image categories, although one of the cheaper algorithms performs nearly as well for some types of distortion. A general hierarchical framework for motion tracking is discussed in [1]. A correlation based matching approach is selected though gradient approaches are also considered.

Despite the age of the NCC algorithm and the existence of more recent techniques that address its various shortcomings, it is probably fair to say that a suitable replacement has not been universally recognized. NCC makes few requirements on the image sequence and has no parameters to be searched by the user. NCC can be used `as is' to provide simple feature tracking, or it can be used as a component of a more sophisticated (possibly multi-resolution) matching scheme that may address scale and rotation invariance, feature updating, and other issues. The choice of the correlation coefficient over alternative matching criteria such as the sum of absolute differences has also been justified as maximum-likelihood estimation.

5 TRANSFORM DOMAIN COMPUTATION

Consider the numerator in (2) and assume that we have images f(x,y)=f(x,y)-f(u,v) and t(x,y)=t(x,y)-t in which the mean value has already been removed:

$$\gamma(u,v) = \sum f(x,y)\, t(x-u,y-v) \quad \ldots\ldots\ldots\ldots\ldots (3)$$

For a search window of size M^2 and a feature of size N^2 (3) requires approximately $N^2 (M - N + 1)^2$ additions and $N^2 (M - N + 1)^2$ multiplications.

Eq. (3) is a convolution of the image with the reversed feature t'(-x,-y) and can be computed by

$$F^{-1}\{F(f)F^*(t)\} \ldots\ldots\ldots (4)$$

where F is the Fourier transform. The complex conjugate accomplishes reversal of the feature via the Fourier transform property F f * (-x) = F *(w).

Implementations of the FFT algorithm generally require that f' and t' be extended with zeros to a common power of two. The complexity of the transform computation (3) is then 12 M^2 log2 M real multiplications and M^2 log2 M real additions/subtractions. When M is much larger than N the complexity of the direct `spatial' computation (3) is approximately N^2M^2 multiplications/additions, and the direct method is faster than the transform method. The transform method becomes relatively more efficient as N approaches M and with larger M, N.

6 FAST CONVOLUTION

There are several well known fast convolution algorithms that do not use transform domain computation [13]. These approaches fall into two categories: algorithms that trade multiplications for additional additions, and approaches that find a lower point on the $O(N^2)$ characteristic of (one-dimensional) convolution by embedding sections of a one-dimensional convolution into separate dimensions of a smaller multidimensional convolution. While faster than direct convolution these algorithms are nevertheless slower than transform domain convolution at moderate sizes [13] and in any case they do not address computation of the denominator of (2).

7 PHASE CORRELATION

Because (4) can be efficiently computed in the transform domain, several transform domain methods of approximating the image energy normalization in (2) have been developed. Variation in the image energy under the template can be reduced by high-pass filtering the image before cross-correlation. This filtering can be conveniently added to the frequency domain processing, but selection of the cutoff frequency is problematic--a low cutoff may leave significant image energy variations, whereas a high cutoff may remove information useful to the match.

8 NORMALIZING

Examining again the numerator of (2), we note that the mean of the feature can be precomputed, leaving

$$\gamma(u,v) = \sum f(x,y)t(x-u,y-v) - f_{u,v} \sum t(x-u,y-\text{.....(5)}$$

Since t' has zero mean and thus zero sum the term $f_{u,v} \sum t(x-u,y-v)$ is also zero, so the numerator of the normalized cross-correlation can be computed using (4).

Examining the denominator of (2), the length of the feature vector can be pre computed in approximately $3N^2$ operations (small compared to the cost of the cross-correlation), and in fact the feature can be pre-normalized to length one. The problematic quantities are those in the expression $\sum_{x,y} [f(x,y)-f_{u,v}]^2$. The image mean and local (RMS) energy must be computed at each u,v, i.e. at $(M-N+1)^2$ locations, resulting in almost $3N^2(M-N+1)^2$ operations (counting add, subtract, multiply as one operation each). This computation is more than is required for the direct computation of (3) and it may considerably out weight the computation indicated by (4) when the transform method is applicable. A more efficient means of computing the image mean and energy under the feature is desired.

These quantities can be efficiently computed from tables containing the integral (running sum) of the image and image square over the search area, i.e.,

$$s(u,v) = f(u,v) + s(u-1,v) + s(u,v-1) - s(u-1,v-1) \quad \&$$

$$s^2(u,v) = f^2(u,v) + s^2(u-1,v) + s^2(u,v-1) - s^2(u-1,v-1) \text{.....(6)}$$

With $s(u,v) = s^2(u,v) = 0$ when either $u,v < 0$. The energy of the image under the feature positioned at u,v is then $e_f(u,v) = s^2(u+N-1,v+N-1)$

$$- \; s^2(u-1,v+N-1)$$

$$- \; s^2(u+N-1,v-1)$$

$$+ \; s^2(u-1,v-1)$$

And similarly for the image sum under the feature.

The problematic quantity $\sum_{x,y} [f(x,y)-f(u,v)]^2$ can now be computed with very few operations since it expands into an expression involving only the image sum and sum squared under the feature. The construction requires approximately $3M^2$ operations, which is less than the cost of computing the numerator by (4) and considerably less than the $3N^2(M-N+1)^2$ required to compute $\sum_{x,y} [f(x,y)-f(u,v)]^2$ at each u,v.

This technique of computing a definite sum from a pre computed running sum has been independently used in a number of fields; a computer graphics application is developed in [5].

The fast algorithm in some cases reduces high-resolution feature tracking from an overnight to an over-lunch procedure. With lower proxy resolution and faster machines, semi-automated feature tracking is tolerable in an interactive system. Certain applications in other fields may also benefit from the algorithm described here.

9 CONCLUSION

The cross correlation is computed based on the transformation, rotation and scaling of images. The cross indexing computes the correlation coefficient and it also increases the magnitude speed up of the Sift descriptors. For larger Images which require more comparison for template matching uses the correlation to identify its similarity between two images. An image with high resolution requires more features to be extracted to identify its similarity. This coefficient helps to compare the difference and it also need more techniques to identify its difference which requires more research for image search. Most systems which uses large scale image search consists of feature tracking approaches to distinguish the similarity of images. This algorithm gives a clear picture of image comparison with the sift detector which can detect the image by the use of feature tracking techniques.

REFERENCES

[1] P. Anandan, ``A Computational Framework and an Algorithm for the Measurement of Visual Motion", Int. Journal of Computer Vision, 2(3), p. 283-310, 1989.

[2] D. I. Barnea, H. F. Silverman, ``A class of algorithms for fast digital image registration", IEEE Trans. Computers, 21, pp. 179-186, 1972.

[3] R. Brunelli and T. Poggio, ``Face Recognition: Features versus Templates", IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042-1052, 1993.

[4] P. J. Burt, C. Yen, X. Xu, ``Local Correlation Measures for Motion Analysis: a Comparative Study'', IEEE Conf. Pattern Recognition Image Processing 1982, pp. 269-274.

[5] F. Crow,``Summed -Area Tables for Texture Mapping", Computer Graphics, vol 18, No. 3, pp. 207-212, 1984.

[6] R. O. Duda and P. E. Hart, Pattern Classification Scene Analysis, New York: Wiley, 1973.

[7] R. C. Gonzalez and R. E. Woods, Digital Image Processing (third edition), Reading, Massachusetts: Addison- Wesley, 1992.

[8] A. Goshtasby, S. H. Gage, and J. F. Bartholic, ``A Two-Stage Cross - Correlation Approach to Template Matching", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 6, no. 3, pp. 374-378, 1984.

[9] C. Kuglin and D. Hines, ``The Phase Correlation Image Alignment Method," Proc. Int. Conf. Cybernetics and Society, 1975, pp. 163-165.

[10] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.

[11] R. C. Gonzalez and R. E. Woods, Digital Image Processing (third edition), Reading, Massachusetts: Addison- Wesley, 1992.