FFT-BASED ESTIMATION OF LARGE MOTIONS IN IMAGES: A ROBUST GRADIENT-BASED APPROACH

Georgios Tzimiropoulos[1]*, Vasileios Argyriou*[2] *and Tania Stathaki*[1]

[1]Imperial College London
London, SW7 2AZ , UK

[2]University of East London
London E16 2RD, UK

ABSTRACT

A fast and robust gradient-based motion estimation technique which operates in the frequency domain is presented. The algorithm combines the natural advantages of a good feature selection offered by gradient-based methods with the robustness and speed provided by FFT-based correlation schemes. Experimentation with real images taken from a popular database showed that, unlike any other Fourier-based techniques, the method was able to estimate translations, arbitrary rotations and scale factors in the range 4-6.

Index Terms— Motion estimation, correlation methods

1. INTRODUCTION

The estimation of relative motions between two images finds applications in a multitude of computer vision tasks such as image registration, video compression and object recognition. In this work, we propose a robust correlation-based scheme which operates in the Fourier domain for the estimation of large translations, rotations and scalings in images.

For the class of similarity transforms, a frequency domain approach to motion estimation possesses several appealing properties. First, through the use of correlation, it enables an exhaustive search for the unknown motion parameters and, therefore, large motions can be recovered with no a priori information (good initial guess). Second, the approach is global which equips the algorithm with robustness to noise. Third, the method is computational efficient. This comes from the *shift property* of the Fourier transform (FT) and the use of FFT routines for the rapid computation of correlations.

The basic principles for motion estimation in the frequency domain were first introduced in [1]. Given two images related by a similarity transform, the relative scaling and rotation affect only the magnitudes of the FTs of the two images. The magnitudes are represented in the log-polar Fourier domain and, scaling and rotation are recovered first using phase correlation [2]. Then, one of the images is scaled and rotated and the residual translation is estimated using phase correlation in the spatial domain. Phase correlation is used instead of standard correlation since it provides much better localization accuracy. Conversion from Cartesian to log-polar is performed using standard interpolation schemes.

The methods in [3],[4] are considered state-of-the art in FFT-based motion estimation. Performance is improved by introducing new sampling schemes which reduce the inaccuracies induced by resampling the magnitude of the FT on the log-polar grid. To recover scalings and rotations, the method in [3] relies on the pseudopolar FT which rapidly computes a discrete FT on a nearly polar grid. The pseudopolar grid serves as an intermediate step for a log-polar Fourier representation which is obtained using nearest-neighbor interpolation. The total accumulated interpolation error is decreased; nevertheless the pseudopolar FFT is not a true polar Fourier representation and the method estimates the rotation and scaling in an iterative fashion. In [4], the log-polar DFT is approximated by interpolating the pseudo-log-polar FT. The method is non-iterative but the gain in accuracy is not significant.

In this work, we provide reasoning, intuition and experimentation which show that accuracy in FFT-based motion estimation depends on the image representation used and the type of correlation employed rather than the method used to approximate the log-polar DFT. In particular, image functions (in both spatial and log-polar domain) are replaced by edge maps which retain both magnitude and orientation information and then gradient-based cross-correlation [5] is performed to estimate the unknown motion parameters. Gradient cross-correlation was originally proposed in the context of subpixel translation estimation. Here, we demonstrate that the merits of a gradient-based approach are fully exploited when large motions are to be considered. Contrary to common belief that FFT-based schemes are unable to handle large motions for real images [6], evaluation with popular image datasets [7] showed that the method was able to estimate scalings in the range 4-6, arbitrary rotations and large translations.

2. ROBUST FFT-BASED MOTION ESTIMATION

Let $I_i(\mathbf{x})$, $\mathbf{x} = [x, y]^T \in \mathcal{R}^2$, $i = 1, 2$ be two image functions. We denote $\widehat{I}_i(\mathbf{k})$, $\mathbf{k} = [k_x, k_y]^T \in \mathcal{R}^2$ the FT of I_i and M_i the

Mr G. Tzimiropoulos' funding for this work is provided by the Systems Engineering for Autonomous Systems (SEAS) Defence Technology Centre established by the UK Ministry of Defence.

magnitude of $\widehat{I}_i$. Polar and log-polar Fourier representations are denoted using $\mathbf{k}_p = [k_r, k_\theta]^T$ and $\mathbf{k}_l = [\log k_r, k_\theta]^T$ respectively, where $k_r = \sqrt{k_x^2 + k_y^2}$ and $k_\theta = \arctan(k_y/k_x)$.

2.1. Translation estimation

Given two images, I_1 and I_2, that are related by an unknown translation $\mathbf{t} = [t_x, t_y]^T \in \mathcal{R}^2$, i.e.

$$I_1(\mathbf{x} + \mathbf{t}) = I_2(\mathbf{x}) \tag{1}$$

$\mathbf{t}$ can be recovered from the $2D$ cross-correlation function $C(\mathbf{u})$, $\mathbf{u}$=$[u, v]^T \in \mathcal{R}^2$ as $\arg_{\mathbf{u}} \max\{C(\mathbf{u})\}$, where

$$C(\mathbf{u}) = I_1(\mathbf{x}) \star I_2(-\mathbf{x}) = \int_{\mathbf{x}} I_1(\mathbf{x}) I_2(\mathbf{x}+\mathbf{u}) d\mathbf{x} \tag{2}$$

From the *convolution theorem* of the FT, C is given by

$$C(\mathbf{u}) = F^{-1}\{\widehat{I}_1(\mathbf{k})\widehat{I}_2^*(\mathbf{k})\} \tag{3}$$

where F^{-1} is the inverse FT and $*$ denotes the complex conjugate operator. The *shift property* of the FT states that if the relation between I_1 and I_2 is given by (1), then, in the frequency domain, it holds

$$\widehat{I}_1(\mathbf{k}) e^{j\mathbf{k}^T\mathbf{t}} = \widehat{I}_2(\mathbf{k}) \tag{4}$$

and therefore (3) becomes

$$C(\mathbf{u}) = F^{-1}\{{M_1}^2(\mathbf{k}) e^{-j\mathbf{k}^T\mathbf{t}}\} \tag{5}$$

The above analysis summarizes the main principles of frequency domain correlation-based translation estimation. For finite discrete images, the FT is efficiently implemented using FFT routines and the algorithm's complexity is $O(N^2 \log N)$, where N is the length of the given images.

In this work, standard $2D$ Cartesian correlation is replaced by gradient cross-correlation (GC) defined as follows

$$GC(\mathbf{u}) = G_1(\mathbf{x}) \star G_2^*(-\mathbf{x}) = \int_{\mathbf{x}} G_1(\mathbf{x}) G_2^*(\mathbf{x}+\mathbf{u}) d\mathbf{x} \tag{6}$$

where

$$G_i = G_{i,x} + jG_{i,y} \tag{7}$$

and $G_{i,x} = \nabla_x I_i$ and $G_{i,y} = \nabla_y I_i$ are the gradients along the horizontal and vertical direction respectively.

2.1.1. Spatial domain analysis

From the definition of GC and using (7), we can easily derive

$$\begin{aligned} GC(\mathbf{u}) &= G_{1,x}(\mathbf{x}) \star G_{2,x}(-\mathbf{x}) + G_{1,y}(\mathbf{x}) \star G_{2,y}(-\mathbf{x}) \\ &+ j\{-G_{1,x}(\mathbf{x}) \star G_{2,y}(-\mathbf{x}) + G_{1,y}(\mathbf{x}) \star G_{2,x}(-\mathbf{x})\} \end{aligned}$$

The imaginary part in the above equation is equal to zero, therefore

$$GC(\mathbf{u}) = G_{1,x}(\mathbf{x}) \star G_{2,x}(-\mathbf{x}) + G_{1,y}(\mathbf{x}) \star G_{2,y}(-\mathbf{x}) \tag{8}$$

Using the polar representation of complex numbers, we define $R_i = \sqrt{G_{i,x}^2 + G_{i,y}^2}$ and $\Phi_i = \arctan G_{i,y}/G_{i,x}$. Based on this representation, (8) can be written as

$$GC(\mathbf{u}) = \int_{\mathbf{x}} R_1(\mathbf{x}) R_2(\mathbf{x}+\mathbf{u}) \cos[\Phi_1(\mathbf{x}) - \Phi_2(\mathbf{x}+\mathbf{u})] d\mathbf{x} \tag{9}$$

Each term in (9) has its own special importance. The magnitudes R_i reward pixel locations with strong edge responses, while the effect of areas of constant intensity level, which do not provide any reference points for motion estimation, is greatly reduced. Therefore only salient structures are considered in the computation of GC. Orientation information is embedded in the cosine kernel. This term is responsible for the dirac-like shape of the GC surface. This can be roughly shown by ignoring the magnitude terms R_i in (9) and making the reasonable assumption that $\Delta\Phi(\mathbf{u}) = \Phi_1(\mathbf{x}) - \Phi_2(\mathbf{x}+\mathbf{u})$ at $\mathbf{u} \neq \mathbf{t}$ is uniformly distributed over $[0, 2\pi)$. Then, $\forall\, \mathbf{u} \neq \mathbf{t}$ the integral in (9) will be equal to zero. This property equips GC with excellent peak localization accuracy.

2.1.2. Frequency domain analysis

From (5), it can be seen that the phase difference term $e^{-j\mathbf{k}^T\mathbf{t}}$, which contains the translational information, is weighted by the magnitude M_1. Then, the inverse FT is taken to yield the standard $2D$ spatial correlation function C. In practise, where (1) holds approximately and $M_1 \neq M_2$, the translational displacement is estimated through (3), and in this case, the phase difference function is weighted by the term $M_1 M_2$. Due to the low pass nature of images, the weighting operation results in a peak of large magnitude in C, however, at the same time, good peak localization is inevitably sacrificed.

GC in the frequency domain is simply defined by replacing $\widehat{I}_i$ with $\widehat{G}_i$ in (3). It can be easily shown that differentiation in the spatial domain is equivalent to high-pass filtering in the Fourier domain. Taking the FT in both parts of (7) yields

$$\widehat{G}_i(\mathbf{k}) = jk_x\widehat{I}_i(\mathbf{k}) - k_y\widehat{I}_i(\mathbf{k}) \tag{10}$$

The magnitude M_{G_i} is given by

$$M_{G_i}(\mathbf{k}) = k_r M_i(\mathbf{k}) \tag{11}$$

and, in this case, the weighting operation results in a peak of large magnitude in GC with very good localization accuracy.

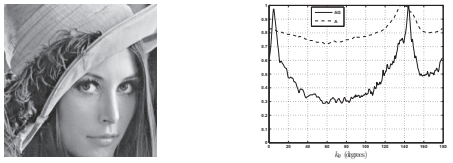
2.2. Estimation of translations, rotations and scalings

Assume now that we are given two images, I_1 and I_2, that are related by a translation $\mathbf{t}$, rotation θ_0 and scaling s, that is

$$I_1(D\mathbf{x} + \mathbf{t}) = I_2(\mathbf{x}) \tag{12}$$

where $D = s \begin{bmatrix} \cos(\theta_0) & \sin(\theta_0) \\ -\sin(\theta_0) & \cos(\theta_0) \end{bmatrix}$. In the Fourier domain, it holds

$$(1/s^2)\widehat{I}_1(D\mathbf{k}/s) e^{j\mathbf{k}^T\mathbf{t}} = \widehat{I}_2(\mathbf{k}) \tag{13}$$



(a) (b)

Fig. 1. (a) "Lena" and (b) the $1D$ representations A (dashed line) and A_G (straight line).

Taking the magnitude in both parts and using the the log-polar representation yields (ignoring the term $1/s^2$)

$$M_1(\mathbf{k}_l - [\log s, \theta_0]^T) = M_2(\mathbf{k}_l) \tag{14}$$

It can be seen that in the log-polar Fourier magnitude domain, scaling and rotation reduce to a $2D$ translation which can be estimated using correlation. After compensating for scaling and rotation, the remaining unknown translation is recovered using correlation in the spatial domain. Note that if $\widetilde{\theta_0}$ is the estimated rotation, then it can be shown that $\widetilde{\theta_0} = \theta_0$ or $\widetilde{\theta_0} = \theta_0 + \pi$. To resolve the ambiguity, one needs to compensate for both possible rotations, compute the correlation functions and, finally, choose as valid solution the one that yields the highest peak [1].

In the proposed scheme, M_1 and M_2 are replaced by M_{G_1} and M_{G_2} and then, after resampling on the log-polar grid, scaling and rotations are estimated using GC. The robustness of the proposed approach is attributed to both the nice properties of GC, as outlined before, and the representation M_{G_i} used as a basis to perform correlation in the log-polar domain. For the latter point, we give three reasons as follows.

First, M_{G_i} naturally emphasizes the frequency bands which reflect the spatial structure of the image salient features. Low frequencies do not provide any reference points for the estimation of θ_0 and s. To illustrate this, consider the "Lena" image and the scenario where the motion is purely rotational and is simply recovered by correlating the $1D$ representation $A(k_\theta) = \int M(k_r, k_\theta) dk_r$ over the angular parameter. The image contains a wide range of frequencies and, consequently, A (Fig. 1 (b), dashed line) is almost flat. In this case, matching by correlation may become unstable. In contrary, A_G (Fig. 1 (b), straight line), obtained by averaging M_G, contains two distinctive peaks which can be used as salient features to perform robust correlation. Additionally, we note that converting from Cartesian to log-polar induces much larger interpolation error for low frequency components. This is because near the origin of the original Cartesian grid less data are available for interpolation. Overall, we conclude that discarding low frequencies from the representation results in better registration accuracy.

Second, using FFT routines to approximate the Fourier spectrum of images results in significant aliasing effects. Rotations and scaling in images induce additional sources of aliasing artifacts which are aggravated by the presence of high frequencies [8]. Through the use of filters, which exhibit band-pass spectral selection properties, to perform differentiation in (7), M_{G_i} is essentially little affected by high-frequency noise and aliasing.

Third, due to the periodic nature of the FFT, in practice, windowing should be applied to the input images to reduce the effect of boundaries whose registration corresponds to zero motion. Making the reasonable assumption that there is no prior knowledge about the motion to be estimated, the same window is typically placed at the center of both images. In this case, windowing not only results in loss of information but also attenuates pixel values in regions shared by the two images in different ways. For large motions, the result is a dramatic decrease in performance. On the other hand, it can be observed that the proposed scheme is based on gradient edge maps and, therefore, discontinuities due to periodization appear only if very strong edges exist close to the image boundaries. In practice, by selecting efficient differentiation operators the boundary effect is greatly reduced and the method does not apply any windowing to the input images.

3. RESULTS

Evaluation was performed using a popular database consisting of 9 different datasets with real images [7]. Each dataset contains a 650x850 reference image and a set of translated, rotated and scaled images of the same resolution. Approximately 90 image pairs, covering a wide range of rotations and scale factors up to 6, were considered.

The proposed scheme was implemented using central differences of second order to perform differentiation, FFT length equal to 1025 to compute M_{G_i}, and bilinear interpolation to obtain the 512x512 log-polar Fourier representations. No additional zero-padding was performed to compute GC in the log-polar domain.

An overview of the results is given in Table 1. For each dataset, we present the maximum scale factor $\widehat{s}$ and the corresponding rotation $\widehat{\theta_0}$ estimated by the algorithm along with the ground truth s and θ_0 as given in [7]. With the exception of "Inria", "Inria Model" and "Ensimag", the algorithm was able to correctly estimate the maximum scale change considered for all datasets. Translations and rotations were estimated to nearly one pixel and degree accuracy respectively. Two examples illustrating the accuracy of registration are given in Fig. 2, where the reference image is scaled, rotated and translated according to the estimated motion parameters, and then superimposed on the target image. To show the gain in performance compared to other FFT-based approaches, we have also implemented an improved version of the state-of-the-art method given in [3]. In particular, the pseudopolar FFT is replaced with an accurate polar FFT [9]. Log-polar Fourier represen-

(a) (b)

Fig. 2. Registration accuracy achieved by the proposed scheme. (a) Ground truth: $(s, \theta_0) = (4.36, 46.0°)$, Estimates: $(\hat{s}, \hat{\theta}_0) = (4.26, 45.7°)$. (b) Ground truth: $(s, \theta_0) = (5.89, 33.2°)$, Estimates: $(\hat{s}, \hat{\theta}_0) = (5.85, 31.6°)$.

	Proposed scheme		Polar FFT	
Dataset	(s, θ_0)	$(\hat{s}, \hat{\theta}_0)$	(s, θ_0)	$(\hat{s}, \hat{\theta}_0)$
"Boat"	$(4.36, 46.0°)$	$(4.26, 45.7°)$	$(1.36, 39.8°)$	$(1.33, 39.6°)$
"East Park"	$(5.77, 0.6°)$	$(5.78, 0.4°)$	$(-)$	$(-)$
"East South"	$(5.09, 60.0°)$	$(5.18, 59.4°)$	$(1.41, 35.6°)$	$(1.37, 35.5°)$
"Inria"	$(4.03, 0.8°)$	$(3.91, 0.7°)$	$(-)$	$(-)$
"Inria Model"	$(4.79, 50.82°)$	$(4.82, 51.0°)$	$(-)$	$(-)$
"Laptop"	$(1.51, 45.4°)$	$(1.51, 45.0°)$	$(-)$	$(-)$
"Resid"	$(5.89, 33.2°)$	$(5.85, 31.6°)$	$(1.12, 32.4°)$	$(1.11, 32.5°)$
"UBC"	$(2.89, 9.6°)$	$(2.89, 9.5°)$	$(1.25, 51.9°)$	$(1.23, 51.9°)$
"Ensimag"	$(4.92, 40.7°)$	$(4.76, 41.5°)$	$(-)$	$(-)$

Table 1. The maximum scale factors and the corresponding rotations recovered by the proposed scheme and the state-of-the-art respectively.

tations are then obtained using bilinear interpolation. The results are given in Table 1. The method failed completely for "East Park", "Inria", "Inria Model", "Laptop" and "Ensimag" datasets and was not able to recover scalings greater than 1.5.

4. DISCUSSION

We have presented a gradient-based approach for FFT-based motion estimation which is able to estimate large motions in real images. The dynamic range of the algorithm generally depends on the application and the image resolution. For the images used in this experiment, the method recovered scalings in the range 4-6 and arbitrary rotations. For most applications, maximum scale changes are not expected to be larger than 3-4. An additional advantage of the proposed approach is that the method is complimentary to other state-of-the-art FFT-based image registration methods. On-going research is focused on performance evaluation of such composite schemes.

5. REFERENCES

[1] B.S. Reddy and B.N. Chatterji, "An fft-based technique for translation, rotation, and scale-invariant image registration," *IEEE Trans. Image Processing*, vol. 5, no. 8, pp. 1266–1271, 1996.

[2] C.D. Kuglin and D.C. Hines, "The phase correlation image alignment method," in *Proc. IEEE Conf. Cybernetics and Society*, 1975, pp. 163–165.

[3] Y. Keller, A. Averbuch, and M. Israeli, "Pseudopolar-based estimation of large translations, rotations and scalings in images," *IEEE Trans. Image Processing*, vol. 14, no. 1, pp. 12–22, 2005.

[4] H. Liu, B. Guo, and Z. Feng, "Pseudo-log-polar fourier transform for image registration," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 17–21, 2006.

[5] V. Argyriou and T. Vlachos, "Performance study of gradient cross-correlation for sub-pixel motion estimation in the frequency domain," *IEE Vis. Image Signal Process.*, vol. 152, no. 1, pp. 107–114, 2005.

[6] G.Wolberg and S. Zokai, "Image registration using log-polar mappings for recovery of large-scale similarity and projective transformations," *IEEE Trans. Image Process.*, vol. 14, no. 10, pp. 1422–1434, 2005.

[7] "http://lear.inrialpes.fr/people/mikolajczyk/," .

[8] H.S. Stone, B. Tao, and M. MacGuire, "Analysis of image registration noise due to rotationally dependent aliasing," *J. Vis. Commun. Image*, vol. R.14, pp. 114–135, 2003.

[9] A. Averbuch, R.R. Coifman, D.L. Donoho, M. Elad, and M.Israeli, "Fast and accurate polar fourier transform," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 145–167, 2006.