MVA2011 IAPR Conference on Machine Vision Applications, June 13-15, 2011, Nara, JAPAN

Spectral-Differential Feature Matching and Clustering for Multi-body Motion Estimation

Anton N. Averkin, Igor P. Gurov, Maxim V. Peterson, Alexey S. Potapov

Saint Petersburg State University of Information Technologies, Mechanics and Optics
49 Kronverksky ave., Saint Petersburg, 197101, Russia
E-mail: ant-averkin@rambler.ru; gurov@mail.ifmo.ru;
maxim.peterson@gmail.com; pas.aicv@gmail.com;

Abstract

System for estimating the motion of independently moving objects observed by a moving camera is presented. It consists of feature matching and multi-body motion estimating modules. Novel set of invariant features is proposed on the base of phase spectrum differentiation without information loss. Clustering the feature points and estimating the transformation model for each cluster are guided by criterion derived from the minimum description length principle that results in correct selection of number of clusters and family of transformations, as well as rejection of outliers.

1. Introduction

Reconstruction of motion models for simultaneously moving camera and several objects in a scene is a problem in the field of machine vision that receives increasing attention. Its solution can be used in different tasks such as mobile robots navigation or augmented reality. This problem can be separated into two tasks: matching the key points, and reconstruction of scene and motion structure.

Existent invariant image representations, which are used in the first task, can be divided into complete and incomplete representations. Complete representations permit reconstruction of initial images with exactness up to transformation, relative to which invariance property is held. Some information about an image is additionally lost in the case of incomplete representations.

Complete invariant representations of images are based on normalization. In particular, some characteristic scale and orientation can be found and then compensated resulting in invariance. This technique is utilized by many widely used features such as SIFT and SURF [1]. Drawback of this approach is its sensitivity to precision of estimated characteristic scale and orientation. Incorrect estimations lead to loss of invariance of local features.

Incomplete representations can be constructed on the base of Fourier-Mellin transform. This approach is less frequent in the task of matching indoor images, but can also be used showing good results [2]. However, such transforms lead to considerable loss of information (e.g. phase information of spatial spectrum in log-polar coordinates) making local features less discriminative.

In this work, we propose novel spectral-differential representation of image as the base of complete invariant representation without use of affine normalization. That is, this representation can be as robust as incomplete representations and as discriminative as complete representations based on normalization. Developed representation is used in the task of key points matching.

There also exist a lot of works devoted to solving the second task. However, the most of previous investigations are based on the assumptions that sequence of images includes motion of a single rigid body, i.e. whether camera moves inside static scene or camera is static, and scene contains only one moving object [3].

More general case of dynamic scene with independently moving objects and camera is of current importance. Algebraic approach to this task exists. This approach is based on generalized principal component analysis (GPCA) [4]. However, these methods are not too robust with respect to outliers occurring because of incorrect matching of key points in a pair of consequent images. Amount of computations also quickly increases with increase of number of independent motion models.

Geometric approach described in the paper [5] groups the key points into clusters corresponding to independently moving objects, which motion can be described using models with different number of parameters. This method utilizes Bayesian approach to model estimation. Root-mean-square or maximum likelihood criteria are not enough, when selection between models with different number of parameters is to be carried out.

In this work, we use the minimum description length principle (MDL [6]) to solve the problem of selection between models with different complexity. This principle has already been applied to the task of motion model selection for a sequence of images in the paper [7], in which displacement of each object was described using essential matrix in assumption of perspective projection. Proposed solution is based on random sampling of key points on a pair of images, which are then distributed between clusters and tracked on the consequent images.

Here, criterion based on the MDL principle is introduced that permits to divide the given array of key points into clusters, each of which is described by own affine or projective transformation and characterizes separate moving object. Novelty consists in clustering algorithm with automatic selection between different motion models and outlier rejection on the base of the MDL principle.

2. Spectral-Differential Representation

Let image $f(x, y)$ be given. Consider its spectrum in continuous case

$$F(\omega_x,\omega_y)=\iint f(x,y)e^{-i(\omega_x x+\omega_y y)}dxdy\ .$$

Spectrum of shifted image $f'(x,y)=f(x–\Delta x,y–\Delta y)$ can be estimated in the form

$$F'(\boldsymbol{\omega})=F(\boldsymbol{\omega})e^{-i(\omega_x\Delta x+\omega_y\Delta y)},\ \ \boldsymbol{\omega}=(\omega_x,\omega_y)\ . \qquad (1)$$

Obviously, $|F'(\boldsymbol{\omega})|=|F(\boldsymbol{\omega})|$ that makes amplitude spectrum convenient tool for invariant image representations. However, a half of information from any image is lost in it. Here, we propose new approach to achieve the same invariance property without loss of phase information.

It can be seen from equation (1) that image displacement results in shifts in harmonic phases proportional to their frequencies ω_x, ω_y and unknown values of displacement Δx, Δy. Spectrum can be expressed in terms of amplitude and phase components: $F(\boldsymbol{\omega})=A(\boldsymbol{\omega})e^{i\varphi(\boldsymbol{\omega})}$.

Spectrum of the shifted image takes the form: $A'(\boldsymbol{\omega})=A(\boldsymbol{\omega})$ and $\varphi'(\boldsymbol{\omega})=\varphi(\boldsymbol{\omega})-\omega_x\Delta x-\omega_y\Delta y$.

Obviously, all the second-order partial derivatives with respect to ω_x and ω_y are invariant relative to shifts.

Any second-order derivative of the phase can be combined with invariable amplitude information in order to get complete shift-invariant representation of images:

$$F_{\text{Invar}_1}(\boldsymbol{\omega})=A(\boldsymbol{\omega})\exp\left[i\frac{\partial^2\varphi(\boldsymbol{\omega})}{\partial\omega_x\partial\omega_y}\right]. \qquad (2)$$

Invariance can be achieved without explicit computation of phases. Consider phase component of spectrum in the form $\Psi(\boldsymbol{\omega})=\exp(i\varphi(\boldsymbol{\omega}))$ for the original image, and $\Psi'(\boldsymbol{\omega})=\exp[i(\varphi(\boldsymbol{\omega})–\omega_x\Delta x–\omega_y\Delta y)]$ for the shifted image. Differencing by ω_x will result in

$$\frac{\partial\Psi'(\boldsymbol{\omega})}{\partial\omega_x}=i\left(\frac{\partial\varphi(\boldsymbol{\omega})}{\partial\omega_x}-\Delta x\right)\exp\left[i\left(\varphi(\boldsymbol{\omega})-\omega_x\Delta x-\omega_y\Delta y\right)\right]\Rightarrow$$

$$\frac{\partial}{\partial\omega_x}\left[\Psi'^{*}(\boldsymbol{\omega})\frac{\partial\Psi'(\boldsymbol{\omega})}{\partial\omega_x}\right]=i\frac{\partial^2\varphi(\boldsymbol{\omega})}{\partial\omega_x^2},$$

where sign * means complex conjugation. Now it can be easily proven that

$$\frac{\partial}{\partial\omega_x}\left[\Psi'^{*}(\boldsymbol{\omega})\frac{\partial\Psi'(\boldsymbol{\omega})}{\partial\omega_x}\right]=\frac{\partial}{\partial\omega_x}\left[\Psi^{*}(\boldsymbol{\omega})\frac{\partial\Psi(\boldsymbol{\omega})}{\partial\omega_x}\right],$$

and derivatives with respect to ω_x can be replaced with derivatives with respect to ω_y. Invariant representation can take form

$$F_{\text{Invar}_2}(\boldsymbol{\omega})=A(\boldsymbol{\omega})\frac{\partial}{\partial\omega_x}\left[\Psi^{*}(\boldsymbol{\omega})\frac{\partial\Psi(\boldsymbol{\omega})}{\partial\omega_x}\right]. \qquad (3)$$

Phase information remains in the proposed spectral-differential invariant representation in contrast to traditional amplitude-spectral invariant representation. A little amount of information is lost because of differencing that makes this representation almost complete. It can be shown that this approach is applicable in discrete case also by replacing differentiation with finite differences.

3. Local Spectral-Differential Features

Well-known Fourier-Mellin transform of the given image consists of the following main steps.

1) Amplitude spectrum $A(\boldsymbol{\omega})$ is calculated. As it was mentioned above, this representation is invariant to shifts. At the same time, scaling and rotation of the initial image results in scaling and rotation of its spectrum.

2) Amplitude spectrum is transformed into log-polar coordinates $A(\rho, \theta)$. Scaling and rotation in Cartesian coordinates become shifting in log-polar coordinates.

3) If one wants to solve the image matching problem, correlation of transformed spectra $A(\rho, \theta)$ and $A'(\rho, \theta)$ of two image can be calculated in order to estimate relative scaling and rotation angle. Then, shifts can be estimated after compensation of found angle and scale factor. If one wants to construct invariant representation, transformed amplitude spectrum $A(\rho, \theta)$ should be subjected to the first operation, i.e. DFT should be applied to $A(\rho, \theta)$ and amplitudes should be kept rejecting phase information. It could be easily seen that the resulting representation will be invariant with respect to similarity group.

Here, we propose to use differentiation of spectra (2, 3) in the Fourier-Mellin transform instead of exclusion of phase information. As the result, one can obtain novel method for matching entire images, and novel invariant representation of images or descriptors of key points.

Since fragments to be described and compared are taken around corresponding key points, centers of these fragments should have approximately zero shifts. Consequently, there is no need to try to achieve invariance to shifts. Indeed, only characteristic scale and rotation (not shifts) are compensated in the methods based on normalization. As the result, our modified Fourier-Mellin transform can be simplified in the context of the given task. The following steps should be performed.

1. Calculating log-polar transform of image around given key point (x_0, y_0) obtaining $f(\rho, \theta\,|\,x_0, y_0)$. This transformation cuts a circle of some radius R=exp(0.5ρ).

2. Applying one of operations of spectrum differentiation obtaining $F_{\text{Invar}}(\omega_\rho, \omega_\theta\,|\,x_0, y_0)$ that gives spectral-differential features (SDF) as key point descriptor. In practice, it is possible to take only a portion of low-frequency harmonics as a descriptor in order to reduce size of descriptor (or feature vector).

Descriptors $F_{\text{Invar}}(\omega_\rho, \omega_\theta\,|\,x_i, y_i)$ and $F'_{\text{Invar}}(\omega_\rho, \omega_\theta\,|\,x_j, y_j)$ of two points from different images can be component-wise compared in order to get similarity measure of key points that can be used in some feature-matching algorithm (i.e. nearest-neighbour). Normalization is necessary, because illumination level can vary.

4. Multiple Motion Model Estimation

Suppose that N key points $\mathbf{x}_i=(x_i, y_i, 1)^{\mathrm{T}}$ on the first image and corresponding N points $\mathbf{x}'_i=(x'_i, y'_i, 1)^{\mathrm{T}}$ on the second image are given. Some set of transformation families $\{T_m(\mathbf{x},\mathbf{p}_m)\}_{m=1}^{M}$ projecting points from the first image into the second image is given, where $\mathbf{p}_m$ is a parameter vector of m-th transformation family, and M is total number of transformation families.

In this paper, we suppose that family, which particular transformation belongs to, is unknown a priori. In addition, coordinates of corresponding points in two images are determined with some errors, and considerable number of correspondences (outliers) can be determined incorrectly. Moreover, the given points should be divided into clusters, affected by independent transformations, and the number of clusters is also unknown.

We consider family of homogenous affine transformation and family described by fundamental matrices. Homogenous affine transformation can be written in a form $\mathbf{x}'=\mathbf{H}_a\mathbf{x}$, where $\mathbf{H}_a$ is the affine matrix with 6 free parameters. In the general case of perspective projection, relation between coordinates of corresponding points in two images can be described with the use of fundamental matrix $\mathbf{F}$. This matrix specifies relation between coordinates of points in the form $\mathbf{x}'^{\mathrm{T}}\mathbf{F}\mathbf{x}=0$ that lay only one constraint in contrast to the affine transformation.

The MDL principle states that such the model should be chosen that minimizes the description length of the data encoded with the model and description length of the model itself. Thus, application of the MDL principle to the task of clustering of identified key points in accordance with their motion models requires estimation of description lengths of clustered points for each transformation family. Relation between coordinates of corresponding points can be written for arbitrary transformation in the form $\mathbf{x}'_i=T_m(\mathbf{x}_i, \mathbf{p}_m)+\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i$ are vectors of deviations ($i=1\ldots N_k$, N_k is the number of key points in the k-th cluster). Deviation vectors should be encoded separately in such the way that one can restore $\mathbf{x}'_i$ from $\mathbf{x}_i$, $\mathbf{p}_m$ and $\boldsymbol{\varepsilon}_i$. That is, we consider description length of coordinates of points in the second image supposing that coordinates of key points in the first image are given.

Consider one k-th cluster of points ($k=1\ldots K$, K is the total number of clusters). In order to describe coordinates of these points on the second image, one needs to describe

– parameters of transformation $\mathbf{p}_m$;

– indices of key points belonging to the cluster;

– deviations $\boldsymbol{\varepsilon}_i$.

Description length of transformation parameters depends on their number. Calculation of affine transformation requires known correspondence of 3 pairs of points that yields 6 equations/parameters, and calculation of fundamental matrix requires known correspondence of 7 pairs of points that yields 7 equations/parameters [3]. Each component of the parameter vector should be described with some number of bits. Optimal precision of this description depends on the number of data elements, which are used in the model construction. Conventional estimation [6] of the n_p component vector parameter $\mathbf{p}$ description length for N data elements is $L_{\mathbf{p}}=0.5n_p\log_2 N_k$.

Indices of key points in the cluster can be described by $L_{ind}=\log_2 C_N^{N_k}$ bits, because there are $C_N^{N_k}$ ways to select cluster consisting from N_k points. Description length of deviations depends on transformation family. We can assume that components of $\boldsymbol{\varepsilon}_i$ are independent and identically distributed random variables in the case of affine transformation. Then, one can calculate description length of deviations as $L_{\boldsymbol{\varepsilon}}=2N_k\log_2(2^{-0.5}\sigma_{|\boldsymbol{\varepsilon}|})$, where $2^{-0.5}\sigma_{|\boldsymbol{\varepsilon}|}$ is estimated mean-square deviation for both $\boldsymbol{\varepsilon}_x$ and $\boldsymbol{\varepsilon}_y$.

Fundamental matrix only tells that point belongs to some epilolar line. In order to encode coordinates of a point, one requires encoding its deviation from epipolar line $\boldsymbol{\varepsilon}_l$, and its position along this line $\boldsymbol{\varepsilon}_a$. Deviations from epipolar line could be small, while positions in corresponding epipolar line could differ very much because of different points depth. Hence, these two displacements have very different distributions. Positions on epipolar lines can be encoded as displacements relative to some average value (e.g. computed via homography) that can also be considered as zero-point for disparity. This zero-point should be included as an additional parameter into the model, and the description length of deviations will take a form $L_{\boldsymbol{\varepsilon}}=N_k(\log_2\sigma_l+\log_2\sigma_a)$.

It can be seen that both families of transformation have now similar forms for deviations description length. Motion models from different families can be chosen using the sum of description lengths $L=L_{\mathbf{p}}+L_{ind}+L_{\boldsymbol{\varepsilon}}$.

To perform clustering one needs to consider profit in description length achieved by utilizing motion models. When key point coordinates are encoded without motion models, each coordinate is encoded with $\log_2 S$ bits, where S is the linear size of images. Total profit for cluster with N_k points will be $(2N_k\log_2 S–L)$ bits. Some points should be included into a cluster if this gives maximum positive profit in the description length.

Overall clustering quality criterion takes form

$$L_{tot}=\sum_{k=1}^{K}\left(2N_k\log_2 S-L_k\right), \tag{4}$$

where L_k is the description length of k-th cluster of points encoded with chosen model. It is necessary to consider profits in description lengths, because there can be many outliers, which are not described by any model.

Now, we can describe clustering algorithm. The task consists in determining such distribution of points in clusters, and constructing such motion models for them, which yield maximum profit in L_{tot}. The following algorithm performs clustering with outlier rejection.

1. Probable correspondences between key points in two images are obtained on the base of described SDF.

2. Minimum number of point correspondences necessary for motion model estimation is taken. Here, 7 random nearby points are taken.

3. Parameters of transformations from each family are estimated. Affine transformation matrix is estimated using linear least squares method, and fundamental matrix is estimated using 7-point algorithm [3].

4. Every point not included into any cluster is considered in order to decide, whether its inclusion into the current cluster leads to increase of profit in the description length or not. Consequently, clusters are expanded with such correspondences of key points that maximize criterion function. Thus, outliers are automatically accounted.

5. Points added to the current cluster are excluded from consideration, and steps 2–4 are repeated, while generation of new cluster leads to profit in description length. Point correspondences not included into clusters at the end of work are assumed to be outliers.

6. Steps 1–5 are performed many times with different random sampling (as in RANSAC algorithms). The best solution is chosen on the base of resulting L_{tot} value.

5. Evaluation

We used pairs of indoor and outdoor images taken from different points of view. In this paper, we didn’t consider the problem of key point detection. Any detection algorithm can be used with almost any invariant feature transform (at least, with both SUFR and SDF). We tried to use as key points both conventional maxima in the difference of Gaussian pyramid and corners and centers of straight lines constructed on the base of extracted contours. Relative matching performance appeared to be similar for different key point detection algorithms.

At first, performance of SDF and SURF was compared independent of multi-body motion model estimation. Nearest-neighbour matching was used in order to compare performance of features in the cleanest way. We considered 20 “difficult” indoor and outdoor image pairs.

It appeared that SDF give slightly better performance in average than SURF: 63% vs. 67% incorrect matches correspondingly. However, different methods work better for different images (see error rates for 5 images in the Table 1). Proposed features showed better performance for images with prominent scale, but they appeared to be more sensitive to displacements in positions of key points. Thus, SDF are promising, but somewhat slower.

Table 1. Error rates for SDF and SURF.

Image pair	Error rate, %	
No.	SDF	SURF
1	64	79
2	68	68
3	61	58
4	69	75
5	57	55

Fig. 1 demonstrates two results of feature matching on image pair using SDF and SURF. One can see that SDF is much less sensitive to scale variations than SURF.

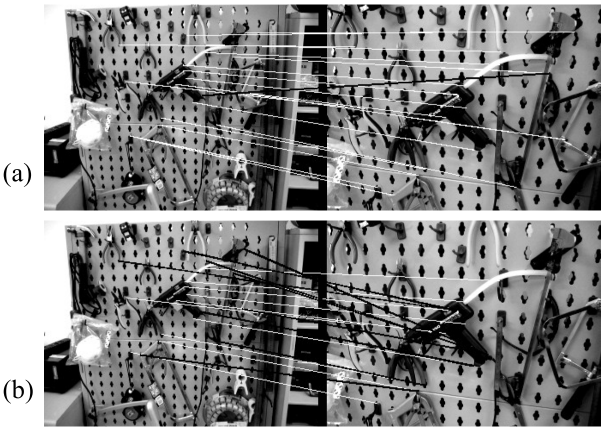

Figure 1. Results of feature matching: using SDF (a), using SURF (b). White lines correspond to correct matches, while black lines correspond to false matches.

Then, we tested the criterion for type of motion model selection separately from clustering problem using images with static scenes. It was confirmed that the deduced MDL criterion permits to select the model with adequate complexity in this task. In particular, selection of less complex model can be seen in Fig. 2. It shows two images taken from rotating camera and found adequate transformation. At that, affine model gives 1085 bit profit, whereas fundamental matrix model gives 1070 bit profit.

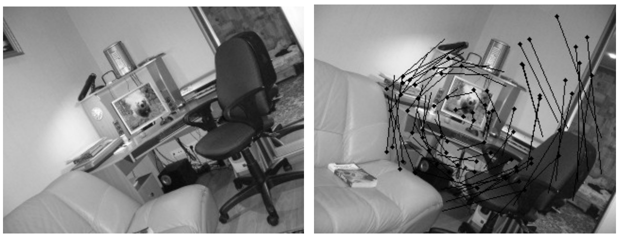

Figure 2. Pair of images and identified key points with displacements described by affine transformation.

Finally, we tested clustering algorithm using images of non-static scenes taken from moving camera. In general, it showed adequate results for both determination of number of motion models and outliers (Fig. 3).

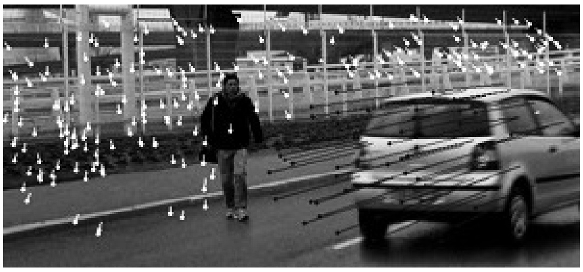

Figure 3. Example of key point clustering.

Conclusions

The task of corresponding key point identification with consequent clustering by means of multi-body motion estimation was addressed. Key point matching was carried out with proposed local invariant feature transform based on modified Fourier-Mellin transform. Modification consisted in replacement of amplitude spectrum calculation with novel operation of phase spectrum differentiation, which doesn't lead to information loss. It was shown that proposed feature transform outperforms SURF in the case of scale changes. Additional research can help to further improve performance of this method.

Criterion of clustering quality was proposed on the base of the MDL principle. This criterion was used to solve the following problems: selection of the adequate number of clusters, selection of family of transformation models with different complexity, and rejection of outliers. Clustering algorithm was developed that optimizes deduced criterion. Adequacy of achieved results was ascertained experimentally. Further investigations should be devoted to the problem of search for the best clustering hypothesis, since random sampling used to avoid local extremes in criterion function could be rather slow.

Work is supported by Russian Federation President's grant Council (MD-2040.2010.9) and Ministry of Education and Science of the Russian Federation.

References

[1] H. Bay, T. Tuytelaars, L. Van Gool: "SURF: Speeded Up Robust Features," *Proc. 9th European Conf. on Computer Vision*. Graz, Austria, vol.3951, pp.404-417, 2006.

[2] A.V. Averkin, A.S. Potapov, V.R. Lutsiv: "Construction of Systems of Local Invariant Image Indicators Based on the Fourier-Mellin Transform," *Journal of Optical Technology*, vol.77, no.1, pp. 28-32, 2010.

[3] R. Hartley, A. Zisserman: "Multiple View Geometry in Computer Vision," *Cambridge University Press*, 2003.

[4] R. Vidal, S. Soatto, Y. Ma, S. Sastry: "Two-view Multibody Structure From Motion," *International Journal of Computer Vision*, vol.68, no.1, pp.7-25, 2006.

[5] P. Torr: "Geometric Motion Segmentation and Model Selection," *Philosophical Transactions of the Royal Society London*, pp.1321-1340, 1998.

[6] J.J. Rissanen: "Modeling by the Shortest Data Description," *Automatica-J.IFAC*, vol.14, pp.465-471, 1978.

[7] K. Schindler, U. James, H. Wang.: "Perspective n-view Multibody Structure-and-Motion through Model Selection," *Proc. ECCV*, pp.606-619, 2006.