A DATA-DRIVEN APPROACH TO FEATURE SPACE SELECTION FOR ROBUST MICRO-ENDOSCOPIC IMAGE RECONSTRUCTION

Pascal Bourdon and David Helbert

XLIM Research Institute (UMR CNRS 7252), University of Poitiers
Bât. SP2MI, Téléport 2, 11 Bd Marie et Pierre Curie, BP 30179
F-86962 Futuroscope Chasseneuil Cedex, France
email: pascal.bourdon@univ-poitiers.fr

ABSTRACT

In the article we propose a new on-line feature space selection strategy for displacement field estimation in the context of multi-view reconstruction of biological images acquired by a multi-photon micro-endoscope. While the high variety of targets encountered in clinical endoscopy induce enough texture feature variability to prohibit the use of recent supervised learning or feature matching-based visual tracking methods, we will show how on-line learning combined with a classical method such as Digital Image Correlation (DIC) can contribute to the improvement of convex optimization-based template matching techniques.

Index Terms— Medical imaging, microendoscopy, optical flow, image mosaicing

1. INTRODUCTION

Matching visual content across images or videos is a major research topic in computer vision, with many applications such as velocimetry, multi-view image reconstruction, or human motion analysis [1, 2, 3, 4, 5, 6]. While many papers focus on supervised learning techniques to address this problem in the context of reproducible scenarios such as face analysis or vehicule tracking, specific applications such as Particle Image Velocimetry (PIV) remain a challenge. A typical example is medical imaging, where strong acquisition noise combined with non-rigid distortions induce feature-point matching systems, such as the well-known SIFT/FLANN association, into errors. As a result, classical techniques based on cross-correlation maximization (e.g. DIC) or mean square error minimization (e.g. Lucas-Kanade [12]) are still very popular, despite certain drawbacks we will discuss later.

In recent years, multi-photon microendoscopy has become an essential tool in cell and tissue biology research [7, 8, 9]. The benefits are its ability to achieve the optical sectioning, high resolution of thick samples (2D image with a low depth of field), high contrast, low sensitivity to the diffusion of biological material, high penetration depth, minimal photo-toxicity and the possibility to dispense with exogenous carcinogen marking. It is another application where the classical DIC method is usually preferred over other methods to estimate displacements prior to the multi-view reconstruction of extracted image data [10].

The displacement estimation method presented in this paper was specifically developed to deal with low field-of-view images acquired with a two-photon micro-endoscope [9]. Such images are obtained by scanning pixel intensity over time along a spiral path. As a result, together with noise, illumination changes, and rigid transformations induced by the hand motion of clinicians operating the endoscope, non-rigid deformations are also generated, mostly in the forms of a fish-eye effect with an additional defective whirl effect in the center of the image, making visual matching a challenging task (see Fig. 1).

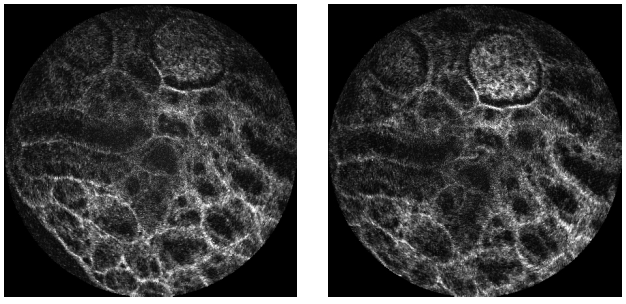

Fig. 1. Consecutive micro-endoscopic acquisitions

While frequency-domain methods such as DIC work very well for rigid transformations, in the future we intend to rely on non-rigid transformation models. As a result, we decided to study model-based methods such as Baker-Matthews' Inverse Computational Image Alignment (ICIA) [11], which is a generalization of Lucas-Kanade [12]. Fur-

The authors would like to thank the XLIM-BIOElectroPhot team for their support, assistance, and most of all for introducing us to this fascinating and challenging subject matter.

thermore, our application requires sub-pixel accuracy, which is natural with ICIA, as opposed to DIC. This however does not mean that DIC is useless to our application. It will be used for ground-truth generation in an on-line learning context, as it is able to address the issue of flow estimation over simple yet large displacements.

The next section of this paper addresses, in terms of the theoretical framework proposed by Baker-Matthews, the issue of image alignment. We will show why the registration of noisy and distorted images cannot always be treated as a conventional convex optimization problem. In section 3, we will discuss the advantages of a feature space approach, and explain how we intend to use DIC results in a feature selection context. Finally, in section 4 we will present experimental results obtained on actual micro-endoscopic image data, and open a discussion about the advantages and drawbacks of this method.

2. VISUAL MATCHING AND CONVEX OPTIMIZATION

2.1. Theoretical framework

The problem of recovering a large field-of-view 2D image projection of a scene using low-field acquisitions can be written as follows : at time t, the endoscope can only capture a projection of said scene on a subset $\Omega \in \Re^2$. Let function I represent this projection for any $\mathbf{x}$ position:

$$\begin{aligned} I: \quad & \Omega \rightarrow \Re \\ & \mathbf{x} \mapsto I(\mathbf{x}, t). \end{aligned} \tag{1}$$

Given a small enough deformation or hand displacement during acquisition, there exists a geometric transform $\mathbf{W}(\mathbf{x}, \theta_\mathbf{t})$ associated with a subset $\Upsilon \in \Omega$ such that:

$$I(\mathbf{x}, t+\partial t) = I(\mathbf{W}(\mathbf{x}, \boldsymbol{\theta}_\mathbf{t}), t) \quad \forall \mathbf{x} \in \Upsilon, \tag{2}$$

where ∂t is the time between two consecutive images, and where $\mathbf{W}(\mathbf{x}, \boldsymbol{\theta})$ denotes a parametrized set of warps that takes a pixel in the coordinate frame of $I(\mathbf{x}, t)$ and maps it to a new, possibly sub-pixel, position in the coordinate frame of $I(\mathbf{x}, t+\partial t)$. $\mathbf{W}(.)$ can be a free-form function or be modelled after rigid or non-rigid transforms.

Subset Υ can either define one or several sets of connected positions, in the case of *block* or *template* matching, or a set of sparse positions, called keypoints, where function I is often replaced by its projection onto a new representation domain called feature space $\mathbf{f}(I)$, in the case of *feature* matching. In all cases, image registration consists in estimating a set of parameters $\boldsymbol{\theta}_\mathbf{t}$ which satisfy Eq. (2), or more likely finding the minimum of the squared residuals:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_\mathbf{t} &= \arg\min(\xi(\boldsymbol{\theta}_\mathbf{t})) \\ \xi(\boldsymbol{\theta}_\mathbf{t}) &= \sum_{\mathbf{x} \in \Upsilon}^{\boldsymbol{\theta}_\mathbf{t}} \| I(\mathbf{x}, t) - I(\mathbf{W}(\mathbf{x}, \boldsymbol{\theta}_\mathbf{t}), t+\partial t) \|_2^2. \end{aligned} \tag{3}$$

2.2. Practical solvability

Objective function $\xi(\boldsymbol{\theta}_\mathbf{t})$ is often considered to be convex and smooth enough for 1^{st} or 2^{nd} order derivatives to exist (either natively or by filtering input images). Therefore Eq. (3) can be solved by taking the partial derivatives of $\xi(\boldsymbol{\theta}_\mathbf{t})$ and setting them equal to zero in a gradient-descent fashion such as the one introduced by Lucas-Kanade:

$$\left. \frac{\partial \xi(\boldsymbol{\theta}_\mathbf{t})}{\partial \boldsymbol{\theta}_\mathbf{t}} \right|_{\boldsymbol{\theta}_\mathbf{t} = \tilde{\boldsymbol{\theta}}_\mathbf{t}} \approx 0, \tag{4}$$

where $\tilde{\boldsymbol{\theta}}_\mathbf{t} \rightarrow \hat{\boldsymbol{\theta}}_\mathbf{t}$ is the expected output.

Fig. 2 shows the result of trying to match a small textured patch between two micro-endoscopic images, using a simple translation model ($\boldsymbol{\theta}_\mathbf{t} \in \Re^2$) such as:

$$\mathbf{W}(\mathbf{x}, \boldsymbol{\theta}_\mathbf{t}) = \mathbf{x} + \boldsymbol{\theta}_\mathbf{t}. \tag{5}$$

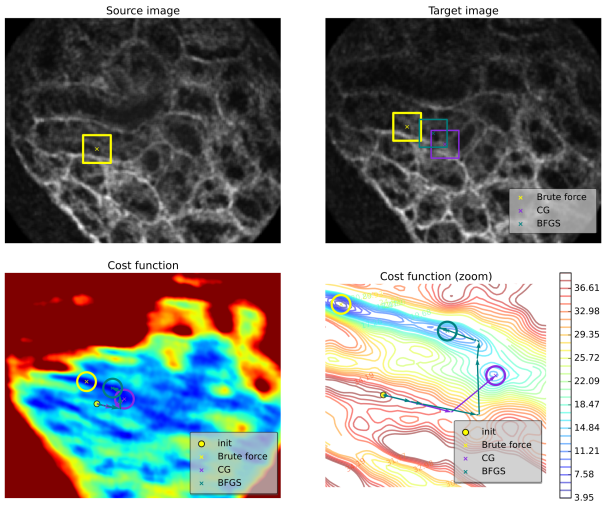

Fig. 2. 2D shift estimation: global, exhaustive (brute-force) search (yellow) succeeds while gradient descent (purple, green) falls into local minima of the cost function.

Two different convex optimization algorithms have been tested and compared to global cross-correlation maximization: the Conjugate Gradient (CG) method, and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. As one can notice on Fig. 2, both methods fail to match the appropriate position in the target image. From a strictly mathematical point of view, it is easy to understand that both methods failed because assumptions made about the convexity of cost function $\xi(\boldsymbol{\theta}_\mathbf{t})$ are inaccurate. Indeed, both CG and BFGS converged into local minima, neither of which are a global minimum. From a visual point of view, we can notice that the two matched image patterns could be good matches if it wasn't for the

pattern matched globally. In other words, the current representation domain of function I is not discriminative enough to distinguish between visually acceptable matches and good, unique matches (e.g. see Zhu *et al.*'s representation model for unsupervised feature selection [13]).

3. DATA-DRIVEN FEATURE SPACE SELECTION

3.1. Feature matching

Feature-matching is often defined as establishing a correspondence between distinct points in images i.e. keypoints [2]. One can however define a feature space where every pixel within the image has a dual representation (e.g. Gabor jets or differences of Gaussians). We define the following vector-valued function:

$$\mathbf{F}(\mathbf{x},t)=\mathbf{f}(I(\mathbf{x},t))=\begin{bmatrix} f_1(I(\mathbf{x},t)) \\ f_2(I(\mathbf{x},t)) \\ \cdots \\ f_N(I(\mathbf{x},t)) \end{bmatrix}, \tag{6}$$

where $\mathbf{F}(\mathbf{x},t)$ is the new representation of function I in a feature space made of a given set of N features $\mathbf{f}(.)$ extracted from image $I(\mathbf{x},t)$. Eq. (3) can be written as:

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_\mathbf{t}(\boldsymbol{\beta})=\arg\min_{\boldsymbol{\theta}_\mathbf{t}}\left(\xi_{\boldsymbol{\beta}}(\boldsymbol{\theta}_\mathbf{t})\right)=\hat{\boldsymbol{\theta}}_\mathbf{t} \\ &\xi_{\boldsymbol{\beta}}(\boldsymbol{\theta}_\mathbf{t})=\sum_{\mathbf{x}\in\Upsilon}\left\|\boldsymbol{\beta}^T.(\mathbf{F}(\mathbf{x},t)-\mathbf{F}(\mathbf{W}(\mathbf{x},\boldsymbol{\theta}_\mathbf{t}),t+\partial t))\right\|_2^2 \end{aligned} \tag{7}$$

with $\boldsymbol{\beta}=\begin{bmatrix}\beta_0 & \beta_1 & \ldots & \beta_N\end{bmatrix}^T$ a vector of weighting coefficients being quantitative expressions of the importance of each element of $\mathbf{f}(.)$, and where $\hat{\boldsymbol{\theta}}_\mathbf{t}$ are the actual *good-match* warp parameters, recovered from exhaustive (brute-force) search or global minimization on the original image domain (Eq. 3).

3.2. On-line space selection

Let $\tilde{\boldsymbol{\theta}}_\mathbf{t}(\boldsymbol{\beta})$ be the set of warping parameters that, given a $\boldsymbol{\beta}$ configuration, provides us with an expected solution for the minimization of $\xi(\boldsymbol{\theta}_\mathbf{t})$. As an example:

$$\left.\frac{\partial\xi_{\boldsymbol{\beta}}(\boldsymbol{\theta}_\mathbf{t})}{\partial\boldsymbol{\theta}_\mathbf{t}}\right|_{\boldsymbol{\theta}_\mathbf{t}=\tilde{\boldsymbol{\theta}}_\mathbf{t}(\boldsymbol{\beta})}\approx 0. \tag{8}$$

Despite DIC not being able to estimate complex warping models, it still provides good results with simple transforms. Therefore we propose to use it for automatic feature selection during a pre-acquisition phase $t\in[0,t_0]$, where the endoscope operator is required to capture a few images for ground truth generation, with DIC running and providing the algorithm with $\hat{\boldsymbol{\theta}}_\mathbf{t}$ samples. While such a phase should be performed whenever substantial changes in the acquisition protocol occur (e.g. when new organs or tissues are targeted), we want to emphasize that this process does not require painful tasks such as manual annotation of ground truth data. The operator only has to capture a few images beforehand and let the algorithm self-calibrate itself. As a matter of fact, given sufficient computational resources, we expect initial DIC-based estimations to perform in real-time, thus providing reconstruction results from the very first captures ($t=0$) before the optimal feature space-based estimator progressively takes over to reduce computer usage.

The estimation of $\hat{\boldsymbol{\beta}}$ is written as:

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\left(\alpha\|\boldsymbol{\beta}\|_1+\sum_{t\in[0,t_0]}\left\|\hat{\boldsymbol{\theta}}_\mathbf{t}-\tilde{\boldsymbol{\theta}}_\mathbf{t}(\boldsymbol{\beta})\right\|_2^2\right), \tag{9}$$

where the ℓ_1-norm constraint on $\boldsymbol{\beta}$ aims towards producing a sparse solution, depending on α values.

Looking for a sparse vector $\hat{\boldsymbol{\beta}}$ has multiple purposes: zeroing elements of $\boldsymbol{\beta}$ reduces the number of features to compute, hence computation time. It also acts as a safety against redundant or non-orthogonal descriptions, reducing the risk of extracting same-information about texture patterns multiple times and having to minimize a non-bijective function in Eq. (9).

4. EXPERIMENTAL RESULTS

The Photonics test data consists of 448 450 × 450 low-field images of a mouse kidney specimen taken with a 250µm fiber endoscope. Ground truth data is generated using global cross-correlation maximization (DIC) to estimate 2D shifts on image patches centered around 205 point positions for each of the first 50 sub-images. Cross-correlation peaks are estimated over full target sub-images in order to easily detect outliers, which are removed from the data set. Eq. (9) is a general ℓ_1-minimization problem of the form $\arg\min_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_1+H(\boldsymbol{\beta})$, where quadratic function $H(\cdot)$ could be assumed as a convex and differentiable function to use an iterative shrinkage solver. However, given the preliminary and experimental nature of this work we decided to limit such assumptions and use the non-realtime stochastic Simulated Annealing (SA) technique for global optimization [14].

The algorithm was tested with two well-known image feature spaces : scale-variant polynomials, using Gaussian (0^{th} order) and Sobel (1^{st} and 2^{nd} orders) kernels of sizes 3×3, 15×15 and 31×31, and Gabor wavelets [15], using 4 equally-spaced orientations for spacial frequencies $\{0.05,0.15\}$ and scales (i.e. σ values) $\{3,7\}$. The BFGS algorithm was used for optical flow estimation on original (raw pixel intensity) and transformed feature spaces.

Table 1 shows results in terms of Mean-Square Error (MSE) and Standard Deviation (SD) computations between estimated displacement fields and ground truth data over the full Photonics sequence. As we can see, our method outperforms intensity-based estimations, with the β-weighted polynomial feature space estimated with $\alpha = 1.0$ providing the best results. It is very interesting to notice that a lower α value of 0.1, which releases the constraint on the ℓ_1-norm of $\boldsymbol{\beta}$, actually leads to worse results than a value of 1.0. We assume this is due to possible non-bijective properties of the cost function discussed earlier, especially since we already know that the initial feature spaces we chose to test are not orthogonal bases.

Feature space	**MSE**	**SD**
Raw pixel intensity	12.67	± 19.40
Multi-scale polynomial, $\alpha = 0.1$	11.05	±16.70
Multi-scale polynomial, $\alpha = 1.0$	**10.37**	**± 14.57**
Gabor wavelet basis, $\alpha = 0.1$	11.21	±16.66
Gabor wavelet basis, $\alpha = 1.0$	11.05	±16.70

Table 1. Estimation results between pixel intensity and β-weighted feature space representations for MSE and SD.

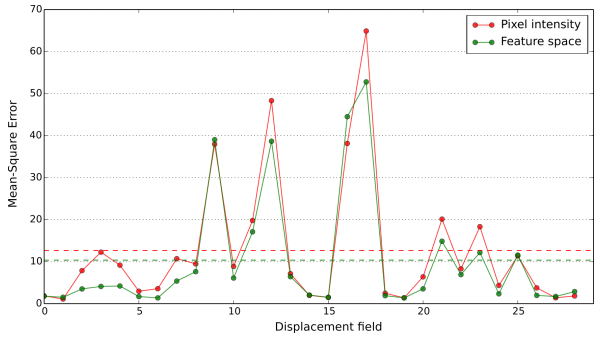

Fig. 3. MSE results between pixel intensity and β-weighted polynomial feature space ($\alpha = 1.0$) for individual displacement fields on a random selection of low-field images.

Both Fig. 3 and Fig. 4 show additional results between pixel intensity and β-weighted, $\alpha = 1.0$, polynomial feature space matching. Fig. 3 shows results for individual field estimations on a random selection of low-field images, while Fig. 4 presents an example of actual estimated displacement fields. Both figures confirm that estimations are improved using our method.

5. CONCLUSION

In this paper we presented a data-driven approach to feature space selection for robust micro-endoscopic image reconstruction. We first discussed how classical template matching techniques such as Digital Image Correlation

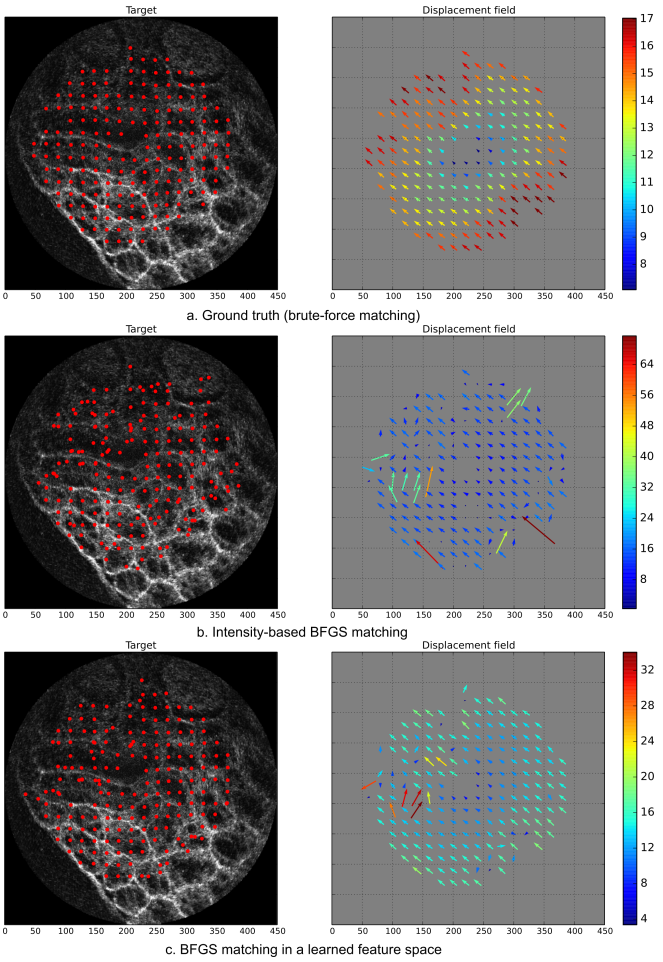

Fig. 4. β-weighted polynomial feature space (c) has more accurate matches than pixel intensity (b).

are still preferred over learned trackers, that are unlikely to succeed with image defects and texture patterns typical of medical images. We also explained how, for the very same reasons, convex optimization-based methods applied in the raw pixel intensity domain can fail, converging into local minima when texture patterns are not unique enough within image data. As a result, we introduced a new feature space selection method to generate template matching-friendly patterns thanks to the on-line learning of ground truth data from non-optimal yet robust global methods such as DIC. Experiments on actual micro-endoscopic images have shown promising results and potential for going further into this research. In future work, we intend to integrate a fully rigid and non-rigid deformation model for global low-field image alignment (as opposed to multiple patch tracking) and improve the on-line learning procedure with a trained detector of data priors such as untrackable points or areas.

6. REFERENCES

[1] J-P Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[2] Adam Baumberg, "Reliable feature matching across widely separated views," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. IEEE, 2000, vol. 1, pp. 774–781.

[3] Michel Riethmuller, Laurent David, and Bertrand Lecordier, *Particle image velocimetry*, Wiley Online Library, 2012.

[4] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios, "Deformable medical image registration: A survey," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[5] William R Crum, Thomas Hartkens, and DLG Hill, "Non-rigid image registration: theory and practice," *The British Journal of Radiology*, 2014.

[6] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1442–1468, 2014.

[7] Noah Bedard, Timothy Quang, Kathleen Schmeler, Rebecca Richards-Kortum, and Tomasz S. Tkaczyk, "Real-time video mosaicing with a high-resolution microendoscope," *Biomed. Opt. Express*, vol. 3, no. 10, pp. 2428–2435, Oct 2012.

[8] Erich E Hoover and Jeff A Squier, "Advances in multiphoton microscopy technology," *Nature photonics*, vol. 7, no. 2, pp. 93–101, 2013.

[9] Guillaume Ducourthial, Pierre Leclerc, Tigran Mansuryan, Marc Fabert, Julien Brevier, Rémi Habert, Flavie Braud, Renaud Batrin, Christine Vever-Bizet, Geneviève Bourg-Heckly, et al., "Development of a real-time flexible multiphoton microendoscope for label-free imaging in a live animal," *Scientific reports*, vol. 5, 2015.

[10] Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup, "Efficient subpixel image registration algorithms," *Optics letters*, vol. 33, no. 2, pp. 156–158, 2008.

[11] Simon Baker and Iain Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[12] Bruce D Lucas, Takeo Kanade, et al., "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, 1981, vol. 81, pp. 674–679.

[13] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon CK Shiu, "Unsupervised feature selection by regularized self-representation," *Pattern Recognition*, vol. 48, no. 2, pp. 438–446, 2015.

[14] A Khachaturyan, S Semenovsovskaya, and B Vainshtein, "The thermodynamic approach to the structure analysis of crystals," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 37, no. 5, pp. 742–754, 1981.

[15] John G Daugman, "Complete discrete 2-d gabor transforms by neural networks for image analysis and compression," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, 1988.