Chapter 6

Image analysis applications

Erkki Oja, Jorma Laaksonen, Jukka Iivarinen, Markus Koskela, Ramūnas Girdziušas, Jussi Pakkanen, Ville Viitaniemi, Mika Rummukainen, Mats Sjöberg

6.1 Content-based image retrieval by self-organizing maps

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Mika Rummukainen, Mats Sjöberg

Content-based image retrieval (CBIR) has been a subject of intensive research effort for more than a decade now. It differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems. In addition, two more points make CBIR systems special. Opposed to such computer vision applications as production quality control systems, operational CBIR systems would be very intimately connected to the people using them. Also, effective CBIR systems call for means of interchanging information concerning images' content between local and remote databases, a characteristic very seldom present, e.g., in industrial computer vision.

PicSOM

The methodological novelty of our neural-network-based CBIR system, PicSOM [1, 2], is to use several Self-Organizing Maps in parallel for retrieving relevant images from a database. These parallel SOMs have been trained with separate data sets obtained from the image data with different feature extraction techniques. The different SOMs and their underlying feature extraction schemes impose different similarity functions on the images. In the PicSOM approach, the system is able to discover those of the parallel SOMs that provide the most valuable information for each individual query instance.

Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [3]. The hierarchical TS-SOM structure is useful for large SOMs in the training phase. In the standard SOM, each model vector has to be compared with the input vector in finding the best-matching unit (BMU). With the TS-SOM one follows the hierarchical structure which reduces the complexity of the search to $O(\log n)$. After training each TS-SOM hierarchical level, that level is fixed and each neural unit on it is given a visual label from the database image nearest to it.

Self-organizing relevance feedback

When we assume that similar images are located near each other on the SOM surfaces, we are motivated to exchange the user-provided relevance information between the SOM units. This is implemented in PicSOM by low-pass filtering the map surfaces. All relevant images are first given equal positive weight inversely proportional to the number of relevant images. Likewise, nonrelevant images receive negative weights that are inversely proportional to their number. The relevance values are then summed in the BMUs of the images and the resulting sparse value fields are low-pass filtered.

Figure 6.1 illustrates how the positive and negative relevance responses, displayed with red and blue map units, respectively, are first mapped on a SOM surface and how the responses are expanded in the low-pass filtering. As shown on the right side of the figure, the relative distances of SOM model vectors can also be taken into account when performing the filtering operation [4]. If the relative distance of two SOM units is small, they can be regarded as belonging to the same cluster and, therefore, the relevance response should easily spread between the neighboring map units. Cluster borders, on the other hand, are characterized by large distances and the spreading of responses should be less intensive.

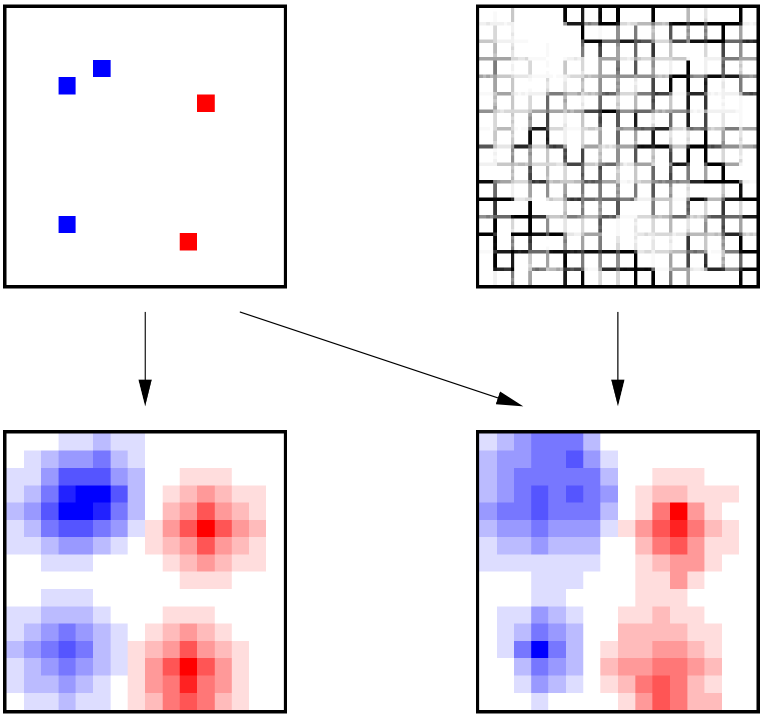


Figure 6.1: An example of how positive and negative map units, shown with red and blue marks on the top-left figure, are low-pass filtered. Two alternative methods exist; either we ignore (bottom-left figure) or take into account (bottom-right figure) the relative distances between neighboring SOM model vectors. In the top-right figure, the relative distances are illustrated with gray level bars so that a darker shade of gray corresponds to a longer relative distance between two neighboring map units.

Finally, the set of images forming the result of the query round is obtained by summing the relevance responses or *qualification values* from all the used SOMs. As a result, the different content descriptors do not need to be explicitly weighted as the system automatically weights their opinions regarding the images' similarity and relevance.

MPEG-7 content descriptors

Development of content-based image retrieval techniques has suffered from the lack of standardized ways for describing visual image content. Fortunately, the MPEG-7 international standard has emerged as both a general framework for content description and a collection of specific, agreed-upon content descriptors. MPEG-7 aims at standardizing the description of multimedia content data. It defines a standard set of descriptors that can be used to describe various types of multimedia information. In the scope of our work, the most relevant part of MPEG-7 is the implementation of a set of still image descriptors. Recently, we have integrated the standard MPEG-7 content descriptors into PicSOM [2] and shown that they can be successfully used with it.

User interaction feature

Relevance feedback can be seen as a form of supervised learning to adjust subsequent query rounds by using information gathered from the user's feedback. It is essential that the learning takes place during one query, and the results are erased when starting a new one. This is because the target of the search usually changes from one query to the next,

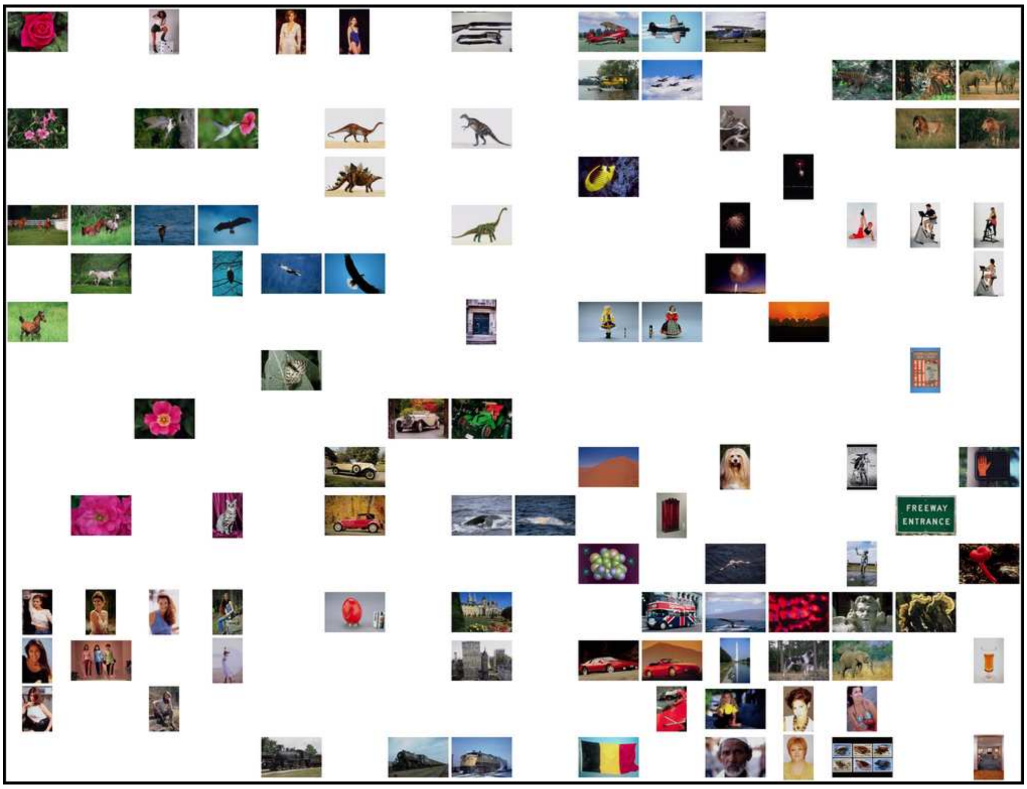


Figure 6.2: The image labels of a 16×16-sized SOM trained with user interaction data.

and so the previous relevances have no significance any more. This is therefore *intra-query* learning.

Relevance feedback provides information which can also be used in an *inter-query* or *long-term* learning scheme. The relevance evaluations provided by the user during a query session partition the set of seen images into relevant and nonrelevant classes with respect to that particular query target. The fact that two images belong to the same class is a cue for similarities in their semantic content. This information can be utilized by considering the previous user interaction as metadata associated with the images and use it to construct a *user interaction* or *relevance feature*, to be used alongside with the visual features. This method was presented and experimented with in [5]. An example of a resulting SOM is illustrated in Figure 6.2. In the figure, a 16×16-sized SOM trained with user interaction data is shown. It can be observed that images with similar semantic content have been mapped near each other on the map.

In some cases, the image database may also contain manually assigned or implicit annotations. These annotations describe high-level semantic content of the image and often contain invaluable information for retrieval purposes. Therefore, it is useful to note that the user-provided relevance evaluations discussed above are notably similar to these annotations. In particular, keyword annotations can be seen as high-quality user assessments and the presented method can be readily utilized also for these annotations.

Use of segmented images

The general problem of image understanding is intrinsically linked to the problem of image segmentation. That is, if one understands an image one can also tell what the different parts of it are. Segmentation thus seems to be a natural part of image understanding, but for an automatic system it is never trivial and the results seldom correspond to the real objects in the picture. But even so segmentation may be useful in CBIR, because different, visually homogeneous regions somehow characterize the objects and scenes in the image. The existing approaches differ mainly in the fashion the segment-wise similarities are combined to form image-wise similarities used in the retrieval.

The implementation of segmentation into PicSOM was done by generalizing the original algorithm so that not only the entire images but also the image segments are seen as objects in their own right. The segments are also considered to be sub-objects of the images they are a part of. The relevance feedback process is modified so that when an image is marked as relevant all its sub-objects (segments) are also marked as relevant. Then, after calculating qualification values for all the objects on the different TS-SOMs, the qualification values of all the sub-objects are summed to their parent objects. Finally, the values obtained from different maps are again summed up to form the final image-wise qualification values.

The results of our preliminary experiments have shown that for most of the used ground truth image classes, the retrieval precision obtained by using both entire and segmented images together excels that obtained by using either ones alone [6]. In a forthcoming series of experiments this will be further ensured.

Application to multimedia messages

We have implemented support for multi-part multimedia messages in the PicSOM system. The system was modified to take advantage of the hierarchical message structure when performing content-based searches. This included implementation of new statistical features for the textual and metadata parts of the messages.

The basic ideas of CBIR can be expanded to more general content-based data retrieval if some kind of low-level statistical feature vectors can be extracted from that particular data. For example, text similarity can be evaluated with the n-gram method. If we also know that some separate data objects are somehow related to each other, we can use this information in the data retrieval. If we, e.g., have a database that contains images of different animals, a short textual description and an audio sample for each of them, we can compare the similarities of the audio samples together with the similarities of the images and texts to obtain the most similar animals as a search result.

We have formed a database of mutually related objects of regular e-mail messages with attachment files, where the message texts and the attachments are probably interrelated. If we now want to search for messages similar to a given reference message, we can use the content-based data retrieval methods of the PicSOM system to first compare similarities of the objects of the same object type in different messages. When these results are then combined, we can evaluate the similarities of the entire messages. On the other hand, instead of searching for whole messages we can use this same approach to help searching for individual message attachments. For example, the text part of a message probably describes what the attachments contain, and the attachments are often related to each other too. If we want to search for an image of a cat from a multimedia message database, we can let the system compare not only the images but also the other related textual objects. The reference message text probably contains the word “cat”, so when searching for images similar to the one in a reference message, we can also compare the texts of the messages and return images attached to the texts containing similar words as the search result.

CBIR benchmarking

The performance of current CBIR systems is still unclear. The lack of common benchmarks or performance measurement methods and standardized ways of communicating have prevented wide-scale performance comparisons. To overcome these difficulties, researchers have been encouraged to make their CBIR systems compatible with the recently developed Multimedia Retrieval Markup Language (MRML), which aims to be the standard for retrieval system communications. Systems communicating with the same language could then easily attend public contests such as the planned Benchathlon contest (*http://www.benchathlon.net*). Currently, there exists one open source CBIR system, the GNU Image Finding Tool (GIFT) developed at the University of Geneva, that uses MRML.

The MRML language has now been implemented into the PicSOM system which is thus able to communicate with other MRML-based applications. The PicSOM system contains a simple method for benchmarking itself. Now, the benchmarking part also communicates via MRML and this has enabled a performance comparison between PicSOM and GIFT to be run [7]. The results based on the recall-relative precision curves were a little surprising—both CBIR systems managed well in some cases, while at the same time both performed badly in other cases. From the seven classes used in the experiment, the PicSOM and the Separate Normalization algorithm of GIFT ranked first for three image classes, while the CIDF algorithm of GIFT performed best for only one class. These results will be valuable when we will continue the development of the PicSOM system.

References

[1] J. T. Laaksonen, J. M. Koskela, S. P. Laakso, and E. Oja. PicSOM - Content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207, November 2000.

[2] J. Laaksonen, M. Koskela, and E. Oja. PicSOM – Self-Organizing Image Retrieval with MPEG-7 Content Descriptors. *IEEE Transactions on Neural Networks*, 13(4): 841-853, July 2002.

[3] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. International Joint Conference on Neural Networks*, vol. II, pages 279-285, Piscataway, NJ, 1990.

[4] M. Koskela, J. Laaksonen, and E. Oja. Implementing Relevance Feedback as Convolutions of Local Neighborhoods on Self-Organizing Maps. In *Proc. International Conference on Artificial Neural Networks*, pages 981-986. Madrid, Spain. August 2002.

[5] M. Koskela and J. Laaksonen. Using Long-Term Learning to Improve Efficiency of Content-Based Image Retrieval. In *Proc. Third International Workshop on Pattern Recognition in Information Systems*, pages 72-79, Angers, France, April 2003.

[6] M. Sjöberg, J. Laaksonen, and V. Viitaniemi. Using Image Segments in PicSOM CBIR System. In *Proc. 13th Scandinavian Conference on Image Analysis*, pages 1106-1113, Halmstad, Sweden, June-July 2003.

[7] M. Rummukainen, J. Laaksonen, and M. Koskela. An efficiency comparison of two content-based image retrieval systems, GIFT and PicSOM. In *Proc. International Conference on Image and Video Retrieval*, pages 500-509, Urbana, IL, July 2003.

6.2 Content-based retrieval of defect images

Jussi Pakkanen, Jukka Iivarinen

A need for efficient and fast methods for content-based image retrieval (CBIR) has increased rapidly during the last decade. The amount of image data that has to be stored, managed, browsed, searched, and retrieved grows continuously on many fields of industry and research.

In this project we have taken a noncommercial CBIR system called PicSOM, and applied it to several databases of surface defect images. PicSOM has been developed in our laboratory at Helsinki University of Technology to be a generic CBIR system for large, unannotated databases. We have made some modifications to the original PicSOM system that affect mostly feature extraction and visualization parts of PicSOM. As an extra problem-specific knowledge we have segmentation masks for each defect image. This information is utilized in PicSOM so that feature extraction is only done for defect areas in each defect image.

Overview of the method

Interpretation of defect images is a demanding task even to an expert. The defect images concerned in this work contain surface defects, and they were taken from a real, online process. Currently we have two major database types: paper and metal surface defects. Both of these types contain several different defect classes (e.g. dark and light spots, holes, scratches, oli stains and so on) that are fuzzy and overlapping, so it is not possible to label defects unambiguously.

In the present work we have adopted the PicSOM system as our content-based image retrieval (CBIR) system and embedded the defect image databases into PicSOM. PicSOM has several features that make it a good choice for our purposes. The most important of these is the fact, that PicSOM can effectively combine search results of different features. This makes adding new features fast and efficient.

Features for defect characterization Several types of features can be used in PicSOM for image querying. These include features for color, shape, texture, and structure description of the image content. When considering defect images, there are two types of features that are of interest: shape features and internal structure features. Shape features are used to capture the essential shape information of defects in order to distinguish between differently shaped defects, e.g. spots and wrinkles. Internal structure features are used to characterize the gray level and textural structure of defects.

One of the advantages of PicSOM is its open architecture. This makes it simple to add new features to the system. Originally we used simple descriptors for shape, texture features based on the co-occurrence matrix, and the gray level histogram. Currently we use the following features, most of which come from the MPEG-7 standard.

Scalable Color descriptor is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform.

Color Layout descriptor specifies a spatial distribution of colors. The image is divided into 8×8 blocks and the dominant colors are solved for each block in the YCbCr color system. Discrete Cosine Transform is applied to the dominant colors in each channel and the DCT coefficients are used as a descriptor.

Color Structure descriptor captures both color content and the structure of this content. It does this by means of a structuring element that is slid over the image. The numbers of positions where the element contains each particular color is recorded and used as a descriptor. As a result, the descriptor can differentiate between images that contain the same amount of a given color but the color is structured differently.

Edge Histogram descriptor represents the spatial distribution of five types of edges in 16 sub-images. The edge types are vertical, horizontal, 45 degree, 135 degree and non-directional, and they are calculated by using 2x2-sized edge detectors for the luminance of the pixels. A local edge histogram with five bins is generated for each sub-image, resulting in a total of 80 histogram bins.

Homogeneous Texture descriptor filters the image with a bank of orientation and scale tuned filters that are modeled using Gabor functions. The first and second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor.

Shape feature For shape description, we use our own problem-specific shape feature set that was developed for surface defect description. It consists of several simple descriptors calculated from a defect's contour. The descriptors are convexity, principal axis ratio, compactness, circular variance, elliptic variance, and angle.

These features were found to work very well on classification experiments using a smaller, pre-classified data base.

Experiments

The problem at hand is now the following one: Given a new defect or a set of defects, retrieve similar defects that might have appeared previously. The retrieval is based on shape and internal structure features, so there is no need for manual annotation or labeling. The largest defect database has almost 45000 defect images that were taken from a real, online process. The images have different kinds of defects, e.g. dark and light spots, holes, and wrinkles. They are automatically segmented beforehand so that each defect image has a gray level image and a binary segmentation mask that indicates defect areas in the image. The image database was provided by our industrial partner, ABB oy.

Two example queries in Figure 6.3 show that the system works quite well. Under the TS-SOMs are the images selected by the user (the so called query images), and at the bottom are the images returned by the PicSOM system. All returned images are visually similar to the query images. The system retains a similar level of success when queried with different types of defects. The true power comes from combining the maps. The PicSOM engine combines the various maps in a powerful manner, yielding good results.

Conclusions

In this project a noncommercial content-based image retrieval (CBIR) system called PicSOM is applied to retrieval of defect images. New feature extraction algorithms for shape and internal structure descriptions are implemented in the PicSOM system. The results of experiments with almost 45000 surface defect images show that the system works fast with good retrieval results.

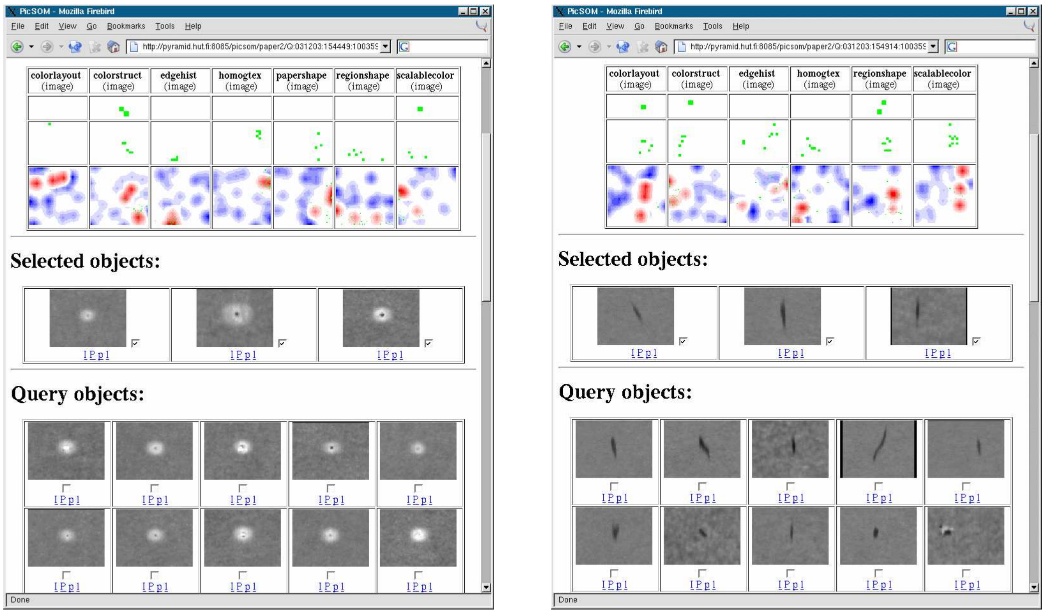


Figure 6.3: Example PicSOM queries.

References

[1] J. Iivarinen and J. Pakkanen, Content-Based Retrieval of Defect Images, In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pp. 62–67, 2002

[2] J. Pakkanen and J. Iivarinen, Content-based retrieval of surface defect images with MPEG-7 descriptors, In K. Tobin Jr. and F. Meriaudeau, editors, *Proceedings of Sixth International Conference on Quality Control by Artificial Vision*, Proc. SPIE 5132, pp. 201–208, 2003.

6.3 Extended fluid-based image registration

Ramūnas Girdziušas, Jorma Laaksonen

Estimation of displacement fields between scalar densities is important in a diverse range of computer vision areas, e.g., patient-to-atlas registration of medical images and object motion field computation.

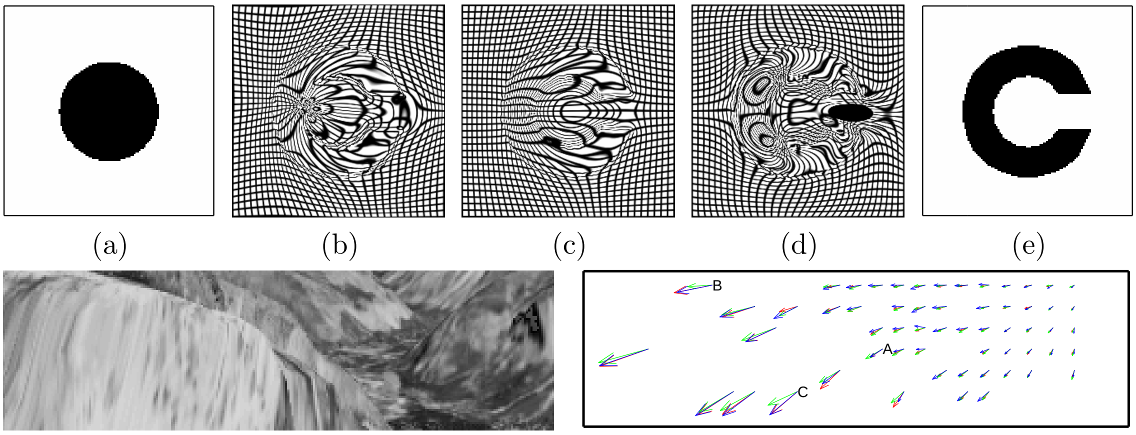
We investigate a family of state-of-the-art fluid-based image registration (FIR) algorithms posed as a constrained optimization problem [1]. In particular, we analyze the Navier-Stokes prior for the velocity field [2], which depends on Lamé constants μ and λ:

$$\int_\Omega \mu||\nabla\mathbf{v} + \nabla\mathbf{v}^T||^2 + 2\lambda(\nabla\cdot\mathbf{v})^2\mathbf{dx}\,. \tag{6.1}$$

The choice of the Lamé constants greatly affects image registration results as can be seen in the toy problem shown in Figure 6.4a-e.

There exists evidence that FIR algorithms perform significantly better, provided that the Lamé constants are enriched with spatio-temporal variability. An example is given in Figure 6.4f-g, where at certain points image intensity-driven fluid-registration algorithm produces 5–10 times lower angular errors of the estimated velocity field than one of the state-of-the-art phase-driven optical flow algorithms.

We are constructing an extended FIR model which: (1) behaves stochastically, (2) allows automatic choice of the Lamé constants, (3) utilizes the information from the image registration results at previous time instants.



(a) (b) (c) (d) (e)

(f) (g)

Figure 6.4: Circle (a) matching to letter ‘C’ (e). Grid deformation is shown: (b) for the weakly elliptic $\lambda + 2\mu \ll 1$ model, (c) Laplacian case $\lambda + \mu = 0$ and (d) strong ellipticity smoothing $\lambda + 2\mu \gg 1$. Optical flow estimation of the simulated ‘fly-through’ Yosemite valley image sequence. Reference frame (f) and optical flow field (g). True flow is depicted in red color, phase-based estimate is shown in green, and fluid-based results in blue.

References

[1] R. Girdziušas and J. Laaksonen, *Multilayer Perceptron Approach to Non-rigid Image Matching.* ICONIP, Vol. 2, pp.491-496, November, 2001.

[2] G. Christensen, *Deformable shape models for anatomy.* PhD thesis, Washington University, August, 1994.