Immediate Structured Visual Search for Medical Images

Karen Simonyan[1], Andrew Zisserman[1], and Antonio Criminisi[2]

[1] University of Oxford, UK
{karen,az}@robots.ox.ac.uk
[2] Microsoft Research, Cambridge, UK
antcrim@microsoft.com

Abstract. The objective of this work is a scalable, *real-time* visual search engine for medical images. In contrast to existing systems that retrieve images that are globally similar to a query image, we enable the user to select a query Region Of Interest (ROI) and automatically detect the corresponding regions within all returned images. This allows the returned images to be ranked on the content of the ROI, rather than the entire image. Our contribution is two-fold: (i) immediate retrieval – the data is appropriately pre-processed so that the search engine returns results in real-time for any query image and ROI; (ii) structured output – returning ROIs with a choice of ranking functions. The retrieval performance is assessed on a number of annotated queries for images from the IRMA X-ray dataset and compared to a baseline.

1 Introduction

The exponential growth of digital medical images of recent years poses both challenges and opportunities. Medical centres now need efficient tools for analysing the plethora of patient images. Myriads of archived scans represent a huge source of data which, if exploited, can inform and improve current clinical practice.

This paper presents a new, scalable, algorithm for the immediate retrieval of medical images *and* structures of interest within them: given a query image and a specified region of interest (ROI) we return images with the corresponding ROI (e.g. the same bone in the hand) delineated. The returned images can be ranked on the contents of the ROI.

Why Immediate Structured Image Search? Given a patient with a condition (e.g. a tumour in the spine) retrieving other generic spine X-rays may not be as useful as returning images of patients with the same pathology, or of exactly the same vertebra. The structured search with an ROI is where we differ from existing content-based medical image retrieval methods which return images that are *globally* similar to a query image [10]. The immediate aspect of our work enables a flexible exploration as it is not necessary to specify in advance what region (e.g. an organ or anomaly), to search for – every region is searchable.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 288–296, 2011.

Use cases include: conducting population studies on specific anatomical structures; tracking the evolution of anomalies efficiently; and finding similar anomalies or pathologies in a particular region. The ranking function can be modified to order the returned images according to the similarity between the query and target ROI's shape or image content. Alternatively, the ROI can be classified, e.g. on whether it contains a particular anomaly such as cysts on the kidney, or arthritis in bones, and ranked by the classification score.

Outline. Sect. 2 describes the retrieval algorithm using X-rays of hands as the running example, with section Sect. 2.4 giving examples of ROI ranking functions. Sect. 3 assesses the retrieval performance and compares to a baseline using images from the publicly available IRMA dataset [3], and Sect. 4 concludes.

Related work and challenges. The problem of ROI retrieval in medical images is addressed in [8], but only in the limiting case of manually pre-segmented ROIs. Our approach is inspired by the Video Google work of [12] for object localisation in videos, and later developments of ROI-driven image retrieval in computer vision for search in large scale image datasets [11]. However, the direct application of these techniques to medical images is not feasible (as shown in Sect. 3) because the feature matching and registration methods of these previous works do not account for inter-subject non-rigid transformations and the repeating structures common to medical images (e.g. phalanx or spine bones). Instead, we employ a registration method tuned to medical images, related to [1]. We differ in that we utilise feature point (landmark) descriptors based on the intensity derivatives, with matching guided to robustly fit non-rigid Thin Plate Spline (TPS) transformations.

2 Structured Image Retrieval Framework

The key to our approach is that registrations are pre-computed (off-line) so that at run time correspondences in target images can be determined immediately for any ROI in the query image. In the case of medical images it is possible to compute registrations between images if they are of the same class, e.g. if they are both images of hands. Given query image and ROI, three stages are involved: (i) image classification, so that ROI correspondences are only considered between images of the same class; (ii) approximate global registration for images within that class, this is pre-computed; and (iii) refinement of the ROI in a target image using the approximate registration as a guide. This is performed at run time. Fig. 1 summarizes the off-line and on-line parts of the framework.

We describe in the following sections how the images are classified; how the global registration is performed; and how the ROI in the target image is refined. Once the ROIs have been accurately localized in each image of the target set, there is then a choice of how these images should be ranked. We describe a number of possibilities for ranking functions in Sect. 2.4.

In the case of a dataset where new images are added, it is important that the method is scalable so that adding new images and using them to make a query

1. **On-line (given a user-specified query image and ROI bounding box)**
   - Select the target image set (repository images of the same class as the query).
   - Using the pre-computed registration (Sect. 2.2), compute the ROIs corresponding to the query ROI in all images of the target set.
   - Refine the ROIs using local search (Sect. 2.2).
   - Rank the ROIs using the similarity measure of choice (Sect. 2.4).
2. **Off-line (pre-processing)**
   - Classify the repository images into a set of pre-defined classes (Sect. 2.1).
   - Compute the registration for all pairs of images of the same class (Sect. 2.2).

Fig. 1. The off-line and on-line parts of the structured retrieval algorithm

does not cause a delay. We have achieved this by registering the query image to only a subset of the same class images; the transformation to the rest can be readily obtained by transform composition as described in Sect. 2.3.

2.1 Image Classification

Our aim is to divide the X-ray images into five classes: `hand, spine, chest, cranium, negative (the rest)`. Certain image retrieval methods take the textual image annotation into account. However, as shown in [5], the error rate of DICOM header information is high, which makes it infeasible to rely on text annotation for classification. Therefore, we use automated visual classification.

We employ the multiple kernel (MKL) technique of [13] and train a set of binary SVM classifiers on multi-scale dense-SIFT and self-similarity visual features in the "one-vs-rest" manner. The classified image is assigned to the class whose classifier outputs the largest decision value. The MKL formulation can exploit different, complementary image representations, leading to high-accuracy classification. The classifier is learnt from training images, and its accuracy is evaluated on a ground truth data set as described in Sect. 3.

2.2 Efficient and Robust Image Registration

In this section we first describe the registration algorithm for a pair of images. This algorithm is the basic workhorse that is used to compute registrations between all images (of the same class). We postpone until Sect. 2.3 how this is done in an efficient and scalable manner. In our case the registration method should be robust to a number of intraclass variabilities of our dataset (e.g. child vs adult hands) as well as additions and deletions (such as overlaid writing, or the wrists not being included). At the same time, it should be reasonably efficient to allow for the fast addition of a new image to the dataset. The method adopted here is a sequence of robust estimations based on sparse feature point matching. The process is initialized by a coarse registration based on matching the first and second order moments of the detected feature points distribution.

This step is feasible since the pairs of images to be registered belong to the same class and similar patterns of detected points can be expected. Given this initial transform T_0, the algorithm then alternates between feature matching (guided by the current transform) and Thin-Plate Spline (TPS) transform estimation (using the current feature matches). These two stages are described next. We use single-scale Harris feature points, and the neighbourhood of each point is described by a SIFT descriptor [9] for matching. Fig. 2 shows examples of the computed registrations.

Guided Feature Matching. Let I_q and I_t be two images to register and T_k the current transform estimate between I_q and I_t. The subscripts i and j indicate matching features in images I_q and I_t with locations $\mathbf{x}_i$, $\mathbf{y}_j$ and descriptor vectors Ψ_i and Ψ_j respectively. Feature point matching is formulated as a linear assignment problem with unary costs $\mathtt{C}_{ij}$ defined as:

$$\mathtt{C}_{ij} = \begin{cases} +\infty & \text{if } \mathtt{C}_{ij}^{geom} > r \\ w^{desc}\mathtt{C}_{ij}^{desc} + w^{geom}\mathtt{C}_{ij}^{geom} & \text{otherwise.} \end{cases} \quad (1)$$

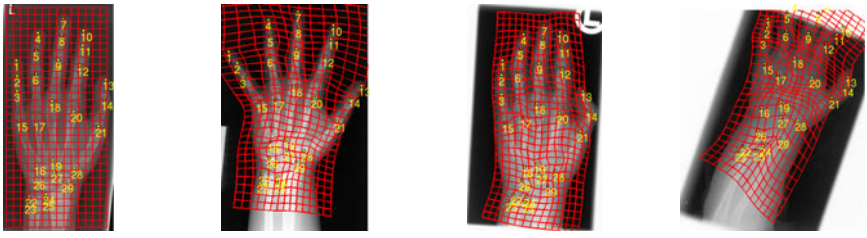
It depends on the descriptors distance $\mathtt{C}_{ij}^{desc} = \|\Psi_i - \Psi_j\|_2$ as well as the symmetric transfer error $\mathtt{C}_{ij}^{geom} = \|T_k(\mathbf{x}_i) - \mathbf{y}_j\|_2 + \|\mathbf{x}_i - T_k^{-1}(\mathbf{y}_j)\|_2$. The hard threshold r on $\mathtt{C}_{ij}^{geom}$ allows matching only within a spatial neighbourhood of a feature. This increases matching robustness while reducing computational complexity.

Robust Thin Plate Spline Estimation. Direct TPS computation based on all feature point matches computed at the previous step leads to inaccuracies due to occasional mismatches. To filter them out we employ the LO-RANSAC [2] framework. In our implementation two transformation models of different complexity are utilised for hypothesis testing. A similarity transform with a *loose* threshold is used for fast initial outlier rejection, while a TPS is fitted only to the inliers of the few promising hypotheses. The resulting TPS warp T_{k+1} is the one with the most inliers.

ROI Localisation Refinement. Given an ROI in the query image we wish to obtain the corresponding ROI in the target, i.e. the ROI covering the same "object". The TPS transform T registering the query and target images provides a rough estimate of the target ROI as a quadrilateral R_t^0 which is a warp of the query rectangle R_q. However, possible inaccuracies in T may cause R_t^0 to be misaligned with the actual ROI, and in turn this may hamper ROI ranking. To alleviate this problem, the detected ROI can be fine-tuned by locally maximizing the normalised intensity cross-correlation between the query rectangle and the target quadrilateral. This task is formulated as a constrained non-linear least squares problem where each vertex is restricted to a box to avoid degeneracies. An example is shown in Fig. 4.

2.3 Scalable Registration by Transform Composition

When adding a new image to the repository, it has to be registered with all target images (images of the same class). A naïve implementation results in the



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

Fig. 2. Robust thin plate spline matching. (a): query image with a rectangular grid and a set of ground-truth (GT) landmarks (shown with yellow numbers); (b)-(d): target images showing the GT points mapped via the *automatically* computed transform (GT points not used) and the induced grid deformation

computational complexity linear in the number of target images, which grows together with the repository size and quickly becomes infeasible as image registration is computationally heavy. Instead, we use a scalable technique suitable for large datasets. A new image q is registered with only a *fixed* subset of E target images (exemplars), which results in E transforms $T_{q,e},\ e = 1 \dots E$. The transformations $T_{e,t}$ between an exemplar e and each of the remaining target set images t are pre-computed. Then the transformation between images q and t can be obtained by transform composition (using different exemplars) followed by robust aggregation as $T_{q,t} = \text{median}_e\,(T_{q,e} \circ T_{e,t})$. While the complexity is still linear in the number of target images, its advantage is that only $E = \text{const}$ registrations should be computed, while transform composition is a cheap operation. The technique is related to the multi-atlas segmentation scheme of [6]. From our experiments (not presented due to the space restrictions), the accuracy of exemplar-based registration is similar to the pairwise case.

2.4 ROI Ranking Functions

At this stage we have obtained ROIs in a set of target images, corresponding to the ROI in the query image. The question then remains of how to order the images for the retrieval system, and this is application dependent. We consider three choices of the ranking function defined as the similarity $S(I_q, R_q, I_t, R_t)$ between the query and target ROIs, R_q, R_t and images I_q, I_t. The retrieval results are ranked in decreasing order of S. The similarity S can be defined to depend on the ROI Appearance (*ROIA*) only. For instance, the normalised cross-correlation (NCC) of ROI intensities can be used. The S function can be readily extended to accommodate the ROI Shape (*ROISA*) as $S = (1 - w)\min(E_q, E_t)/\max(E_q, E_t) + w\,\text{NCC}(R_q, R_t)$, where E_q and E_t are elongation coefficients (ratio of major to minor axis) of query and target ROIs, and $w \in [0, 1]$ is a user tunable parameter. At the other extreme, the function S can be tuned to capture global Image Geometry (*IG*) cues. If similar scale scans are of interest, then S can be defined as: $S(I_q, R_q, I_t, R_t) = (1 - w)\min\{\Sigma, 1/\Sigma\} +$

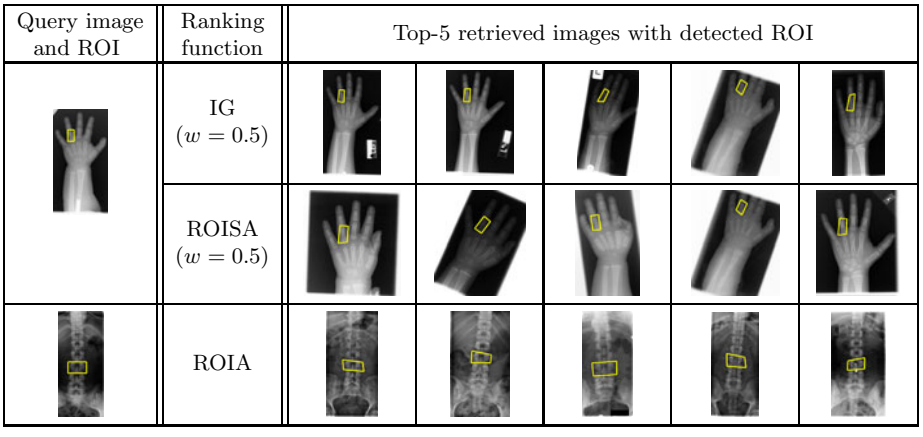


Fig. 3. The effect of different ranking functions on ROI retrieval. ROIs are shown in yellow. IG retrieves scans with similar image cropping; ROISA ranks paediatric hands high because the query is paediatric; ROIA ranks based on ROI intensity similarity.

$w\,\mathrm{NCC}(R_q, R_t)$, where $\Sigma > 0$ is the scale of the similarity transform computed from feature point matches, and $w \in [0, 1]$ is a user tunable parameter.

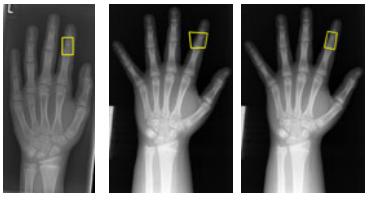
Fig. 3 shows the top ranked images retrieved by these functions. This is an example of how local ROI clues can be employed for ranking, which is not possible with global, image-level visual search. In clinical practice, ranking functions specifically tuned for a particular application could be used, e.g. trained to rank on the presence of a specific anomaly (e.g. nodules or cysts).

3 Results and Comparisons

The dataset. The dataset contains X-ray images of five classes: `hand`, `spine`, `chest`, `cranium`, `negative (the rest)` taken from the publicly available IRMA dataset [3]. Each class is represented by 205 images. The negative class contains images of miscellaneous body parts not included in the other classes. The images are stored in the PNG format without any additional textual metadata. Images within each class exhibit a high amount of variance, e.g. scale changes, missing parts, new objects added (overlaid writings), anatomy configuration changes (phalanges apart or close to each other). Each of the classes is randomly split into 65 testing, 70 training, and 70 validation images.

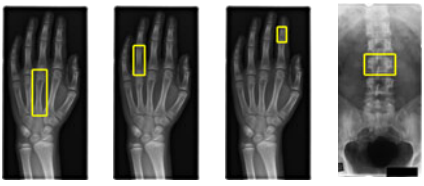
Image classification performance is measured by the ratio of correctly classified test images to the total number of test images. The overall accuracy is 98%. The few misclassifications are caused by the overlap between the `negative` class and other classes, if the negative image partially contains the same body part.

Accuracy of Structured Image Retrieval. To evaluate the accuracy of ROI retrieval from the dataset, we annotated test hand and spine images with



(a) (b) (c)

Fig. 4. (a): query; (b),(c): target before and after local refinement



hand1 hand2 hand3 spine1

Fig. 5. Four annotated bones used for the retrieval performance assessment

Table 1. Comparison of image retrieval accuracy

Method	hand1		hand2		hand3		spine1	
	meanAP	medAP	meanAP	medAP	meanAP	medAP	meanAP	medAP
Proposed	**0.81**	**0.89**	**0.85**	**0.90**	**0.65**	**0.71**	**0.49**	**0.51**
Baseline	0.68	0.71	0.66	0.71	0.38	0.36	0.35	0.35
`elastix`	0.62	0.67	0.61	0.68	0.38	0.37	0.22	0.19

axis-aligned bounding boxes around the same bones as shown in Fig. 5. The ROI retrieval evaluation procedure is based on that of PASCAL VOC detection challenge [4]. A query image and ROI are selected from the test set and the corresponding ROIs are retrieved from the rest of the test set using the proposed algorithm. A detected ROI quadrangle is labelled as correct if the overlap score between its axis-aligned bounding box and the ground truth one is above a threshold. The retrieval performance for a query is assessed using the Average Precision (AP) measure computed as the area under the "precision vs recall" curve. Once the retrieval performance is estimated for each of the images as a query, its mean (meanAP) and median (medAP) over all queries are taken as measures. We compare the retrieval performance of the framework (ROIA ranking, no ROI refinement) using different registration methods: the proposed one (Sect. 2.2), baseline feature matching with affine transform [11], and `elastix` B-splines [7]. All three methods compute pairwise registration (i.e. no exemplars).

The proposed algorithm outperforms the others on all types of queries (Table 1). As opposed to the baseline, our framework can capture non-rigid transforms; intensity-based non-rigid `elastix` registration is not robust enough to cope with the diverse test set. Compared to hand images, worse performance on the spine is caused by less consistent feature detections on cluttered images.

Complexity and Speed. The retrieval framework is efficient and scalable, which allows for the immediate ROI retrieval (in a fraction of second) even using our current non-optimised Matlab implementation. Image class information as well as pairwise registrations between same-class images are pre-computed and stored for the repository images, so they are immediately available if a query image is from the repository. The complexity of adding a new image splits into the following parts: (i) image classification has constant complexity (≈ 0.5s per

image); (ii) registration with a *fixed* number of target set exemplars has constant complexity ($\approx$ 2s per exemplar); (iii) registration with the rest of the target images is linear in the number of images, but the transform composition is very quick compared to registration. Therefore, the most computationally complex operations are invoked only a fixed number of times.

4 Conclusion

We have presented a new framework for the immediate retrieval of medical images and simultaneous, automatic localisation of anatomical structures of interest. Robustness with respect to repeated structures is incorporated via non-rigid image registration driven by guided robust sparse feature matching. Supervised image-level classification also contributes to the high level of accuracy demonstrated on a publicly available labelled database of X-ray images. The proposed visual search framework is fairly generic and can be extended to different modalities/dimensionalities with a proper choice of intra-class registration methods. An interactive demo of the ROI retrieval framework is available on `http://www.robots.ox.ac.uk/~vgg/research/med_search/`

References

1. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Comp. Vis. and Image Understanding 89(2-3), 114–141 (2003)
2. Chum, O., Matas, J., Obdržálek, Š.: Enhancing RANSAC by generalized model optimization. In: Proc. of Asian Conf. on Comp. Vis., vol. 2, pp. 812–817 (2004)
3. Deserno, T.M.: IRMA dataset (2009), `http://ganymed.imib.rwth-aachen.de/irma/datasets_en.php?SELECTED=00009#00009.dataset`
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. Int. J. Comp. Vis. 88(2), 303–338 (2010)
5. Guld, M.O., Kohnen, M., Keysers, D., Schubert, H., Wein, B., Bredno, J., Lehmann, T.M.: Quality of DICOM header information for image categorization. In: Int. Symp. Med. Imag. SPIE, vol. 4685, pp. 280–287 (February 2002)
6. Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B.: Multi-atlas-based segmentation with local decision fusion – application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imag. 28(7), 1000–1010 (2009)
7. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: elastix: A toolbox for intensity-based medical image registration. IEEE Trans. Med. Imag. 29(1), 196–205 (2010)
8. Lam, M., Disney, T., Pham, M., Raicu, D., Furst, J., Susomboon, R.: Content-based image retrieval for pulmonary computed tomography nodule images. In: Med. Imag. 2007: PACS and Imag. Inform. SPIE, vol. 6516 (March 2007)
9. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comp. Vis. 60, 91–110 (2004)

10. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. Int. J. Med. Inform. 73(1), 1–23 (2004)
11. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. IEEE Conf. on Comp. Vis. and Patt. Recog. (2007)
12. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Trans. Patt. Anal. and Mach. Intell. 31(4), 591–606 (2009)
13. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proc. Int. Conf. Comp. Vis., pp. 606–613 (2009)