# Unsupervised Identification of Clinically Relevant Clusters in Routine Imaging Data

Johannes Hofmanninger, Markus Krenn, Markus Holzer, Thomas Schlegl, Helmut Prosch, and Georg Langs *

Department of Biomedical imaging and Image-guided Therapy
Computational Imaging Research Lab, Medical University of Vienna, Austria
johannes.hofmanninger@meduniwien.ac.at, www.cir.meduniwien.ac.at

**Abstract.** A key question in learning from clinical routine imaging data is whether we can identify coherent patterns that re-occur across a population, and at the same time are linked to clinically relevant patient parameters. Here, we present a feature learning and clustering approach that groups 3D imaging data based on visual features at corresponding anatomical regions extracted from clinical routine imaging data without any supervision. On a set of 7812 routine lung computed tomography volumes, we show that the clustering results in a grouping linked to terms in radiology reports which were not used for clustering. We evaluate different visual features in light of their ability to identify groups of images with consistent reported findings.

## 1 Introduction

The number of images produced in radiology departments is rising rapidly, generating thousands of records per day that cover a wide range of diseases and treatment paths [9]. Identifying diagnostically relevant markers in this data is a key to improving diagnosis and prognosis. Currently, computational image analysis typically relies on well annotated and curated training data such as COPDGene or LTRC[1] that have fostered substantial methodological advance. While these kind of data sets enable the creation of accurate and sensitive detectors for specific findings, they are limited, since annotation is only feasible on a relatively small number of cases. Selection or study specific data acquisition can introduce bias, and limits the range of observations represented in the data. In contrast, learning from routine data could enable the discovery of relationships and markers beyond those that can be feasibly annotated, sampling a wide variety of cases. Furthermore, unsupervised learning on such data enables the search for novel disease phenotypes that better reflect a grouping of patients with similar prognosis, than current categories do.

* This research was supported by teamplay which is a Digital Health Service of Siemens Healthineers, by the Austrian Science Fund, FWF I2714-B31, and WWTF S14-069.

[1] www.copdgene.org (COPDgene), ltrcpublic.com (LTRC)

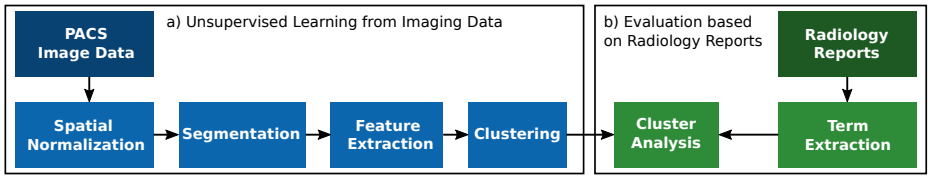


**Fig. 1. Population Clustering & Evaluation**. (a) All processing steps towards population clustering are performed unsupervised and use anonymized routine images exported from a PACS system. (b) Findings extracted from radiology reports are used to evaluate if clusters reflect disease phenotypes in the population.

In this paper, we propose unsupervised learning to group patients based on non-annotated clinical routine imaging data. We show that based on learned visual features, we identify population clusters with homogeneous (within clusters) but distinct (across clusters) clinical findings. To evaluate the link between visual clusters and clinical findings, we compare clusters with corresponding radiology report information extracted with natural language processing algorithms. An overview of the workflow is given in Fig. 1.

***Relation to previous work*** Radiomics [11] involving (a) imaging data, (b) segmentation, (c) feature extraction and (d) analysis [10] has recently gained significant attention, but approaches that reduce the reliance on annotation to extend the covering of variability are scarse. Our work is a contribution to this direction. Although applicable to a large number of conditions, radiomics is mostly applied and developed in oncology [1, 3, 11]. Aerts et al. use a large number of routine CT images of cancer patients recorded on multiple sites to discover prognostic tumor phenotypes [1]. Wibmer et al. differentiate malign from benign prostate tissue by analysing texture features extracted from MRI images [17]. Shin et al. learn semantic associations between radiology images and reports from a data set extracted from a PACS [14], but only uses pre-selected 2D key slices that were referenced from clinicians.

The proposed radiomics approach differs from previous techniques in several significant aspects. We do not restrict analysis to a certain disease type or a small region of interest but implement a general form of population analysis. The most significant difference to prior work is that human interaction is not a prerequisite to bring images into processable form. We do not require selection of key images [14] or manual annotation of regions of interest [1, 11, 17]. In order to make this possible, spatial normalization involving localization and registration is performed. The resulting non-linear mapping to a common reference space allows coordinates and label-masks to be transferred across the population. We extract texture and shape features and use Latent Dirichlet Allocation (LDA) [2] to discover latent topics of co-occurring feature classes that are shared across the population. Subsequently, these topics are used to build volume descriptors by encoding the contribution of each topic to a specific subject.

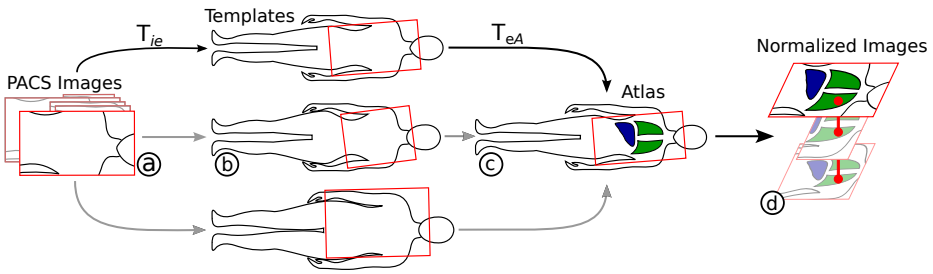


**Fig. 2. Multi-Template Approach** During normalization, an image (a) is aligned to multiple templates (b). All templates are aligned with the atlas (c) by a high quality registration. An image is mapped to the atlas by concatenation of the two corresponding transformations that yield maximal registration quality. After normalization, coordinates and label masks are mapped across the population (d).

## 2 Identification of clusters

***Spatial Normalization*** We perform spatial normalization to establish spatial correspondences of voxels across the population. This allows to study location dependent visual variation without the need for manual definition of regions of interest or preselection of imaging data only showing a specific organ. For this purpose, we perform non-linear registrations of all images to a common reference atlas. For a given image $\mathbf{I}_i \in \{\mathbf{I}_1, \ldots, \mathbf{I}_I\}$ and an atlas $\mathbf{A}$, we seek to find a non-linear transformation $\mathbf{T}$ so that $\mathbf{A} \approx \mathbf{T}(\mathbf{I}_i)$. High variability in the data such as the absence of organs, variation in size and shape or diseases poses challenges to such a registration process. To consider parts of this variations in the normalization process we implement a multi-template approach (Fig. 2). Instead of a direct mapping to an atlas, images are registered to a set of template candidates $\{\mathbf{E}_1, \ldots, \mathbf{E}_E\}$ that cover variability in the population. The transformations of the templates to the atlas are performed in advance, when building the template-set. They are carefully supervised and supported by manually annotated landmarks to ensure high quality registrations.

Let $\mathbf{T}_{ie}$ denote a non-linear transformation from $\mathbf{I}_i$ to a template $\mathbf{E}_e$ and $\mathbf{T}_{eA}$ the transformation from $\mathbf{E}_e$ to $\mathbf{A}$. $\mathbf{I}_i$ is then mapped to $\mathbf{A}$ by concatenating both non-linear transformations so that $\mathbf{A} \approx \mathbf{T}_{eA}(\mathbf{T}_{ie}(\mathbf{I}_i))$. The use of multiple templates gives a candidate set of registrations of a fragment to the reference atlas. Normalized Cross Correlation (NCC) is then used as a quality criteria to select the best transformation by

$$\arg\max_{1 \leq e \leq E} NCC(\mathbf{A}, \mathbf{T}_{eA}(\mathbf{T}_{ie}(\mathbf{I}_i))) \tag{1}$$

In most cases, radiology images cover a delimited region rather than the whole body. To identify location and extend of these fragments in the templates, we perform rigid and affine transformations. An initial rigid position estimation is performed by utilizing correlated 3D-SIFT features [15]. For the template set and

the atlas, we use 17 volumes of the VISCERAL Anatomy 3 dataset [4], which provides CT volumes paired with manually annotated landmarks and organ masks. Non-linear registrations are performed on an isotropic voxel resolution of 2mm using Ezys [5].

***Feature Extraction*** We extract two types of features that capture complementary visual characteristics in order to map an image to a visual descriptor representation so that $\mathbf{I}_i \mapsto \mathbf{f}_i$.
***1. Texture Features*** We densely sample Haralick [6] features of orientation independent Gray-Level Co-occurrence Matrices similar to the work in [16]. Haralick features are able to encode 3D texture and have been used to classify lung diseases [7, 16] or distinguish between cancerous and benign breast tissue [17].
***2. Shape Features*** We extract 3D-SIFT [15] features to encode rotation variant gradient changes such as shape. 3D-SIFT has been used in diagnosis of lung and brain diseases [8, 13].
***3. Bag of Words*** We follow the *Bag Of Visual Words* paradigm to summarize local features to global volume descriptors. In advance, we augment the features with their spatial position in the reference space. This enables to train spatio-visual vocabularies. To account for the different occurrence frequencies of small and large 3DSIFT features, we train two separate vocabularies, microSIFT (3D-SIFT features with $\leq 2cm$ in diameter) and macroSIFT (diameter $> 2cm$). We denote $\mathbf{f}_i^H$ (Haralick) and $\mathbf{f}_i^S$ (SIFT) as the word count feature representations for an image $\mathbf{I}_i$.
***4. Embedding*** Finally, we learn a set of 20 latent topics of co-occurring feature settings of $\mathbf{f}^H$ and $\mathbf{f}^S$ using *Latent Dirichlet Allocation* (LDA) [2]. This allows to interpret an image as a mixture of topics represented by its 20 dimensional topic assignment vector $\mathbf{f}_i^L$.

***Clustering*** We perform clustering of the population to retrieve groups of subjects with (visually) similar properties. Here we interpret the Euclidean distance between two volume descriptors as a measure of visual similarity. This allows to extract clusters utilizing vector quantization. We use the k-means algorithm, an iterative procedure to minimize the sum of squared errors within the clusters, for this purpose. Each subject $i$ is mapped to one clusters $c(i) : i \mapsto k, k \in \{1, \ldots, K\}$. We evaluate if these clusters based on visual information reflect homogeneous profiles of findings in the corresponding radiology reports.

## 3 Evaluation

***Data*** Experiments are performed on a set of 7812 daily routine CT scans acquired in the radiology department of a hospital. The dataset includes all CT scans that were taken during a period of 2 1/2 years and show the lung. We only include volumes with slice thickness of $\leq 3mm$, where the number of slices exceeds 100 and a high spatial frequency reconstruction kernel (e.g. B60, B70, B80, I70, I80,...) was used. For a subset of 5886 cases, the radiology reports in the form of unstructured text are available.

***Term extraction*** We build a NLP framework for automatic extraction of terms describing pathological findings in radiology reports. Extracted terms are mapped to the RadLex[2] ontology, which provides a unified vocabulary of clinical terms, and models relationships by mapping into multiple hierarchies. One of these hierarchies comprises all words that are related to pathological findings. We identify pathological terms by searching for words and their synonyms in the report that are part of this specific hierarchy. The words are then mapped to their respective RadLex term. Our framework is furthermore able to identify negations, so that explicitly negated terms are ignored. We define $T$ as the number of distinct pathological terms and substitute each term by an integer number $\{1, \ldots, T\}$. We define $\mathcal{T}_i$ as the set of all terms that occur in the radiology report of subject $i$. For further analysis we only consider terms that occur more than 50 times resulting in a set of $T = 69$ distinct terms.

***Evaluating associations between visual clusters and report terms*** For evaluation, we restrict the area of interest to the lung, so that only features extracted in the lung are used. Clustering is performed on the full set of images, while for evaluation only records with a report are considered. Aim of the evaluation is to test the hypothesis, that the clustering reflects pathological subgroups in the population. In order to do so we test whether volume label assignments (pathology terms) are associated with cluster assignments. A cell-$\chi^2$-test is performed for each term $t \in \{1, \ldots, T\}$ and each cluster $k \in \{1, \ldots, K\}$ to test whether its cluster frequency $V$ is significantly different from its population frequency $C$ by a $2 \times 2$ contingency table:

| | Term | | |
|---|---|---|---|
| Cluster | present | not present | Total by Cluster |
| In Cluster | $V$ | $R - V$ | $R$ |
| Not in Cluster | $C - V$ | $B - R - C + V$ | $B - R$ |
| Total by Terms | $C$ | $B - C$ | $B$ |

$B$ = Total number of subjects
$V = \sum_{i=1}^{B} [c(i) = k \wedge t \in \mathcal{T}_i]$
$R = \sum_{i=1}^{B} [c(i) = k]$
$C = \sum_{i=1}^{B} [t \in \mathcal{T}_i]$

Here, $B$ denotes the total number of subjects in the population and $R$ the size of a cluster. Since $V$ is potentially small, we perform Fisher's exact test. This results in a p-value that gives the statistical significance of term $t$ being over or under represented in cluster $k$. Testing for each cluster independently increases the Family-Wise Error (FWE) rate and inflates the probability of making a false discovery of an association between the term and a cluster. We strongly control the FWE by correcting the p-values with the Bonferroni-Holmes approach. We define $p_{tk}$ as the corrected p-value for term $t$ being associated with cluster $k$ and $OR_{tk}$ as the corresponding Odds Ratio. As this is an exploratory analysis we do not correct the p-values on the term level.

***Quality Criterion of Clusters*** We interpret the number of discovered associations between cluster and terms as a measure of quality of the population clustering. This not only allows to quantify the relative quality of an image descriptor, but also enables to find the optimal number of clusters. For a predefined

[2] http://www.rsna.org/RadLex

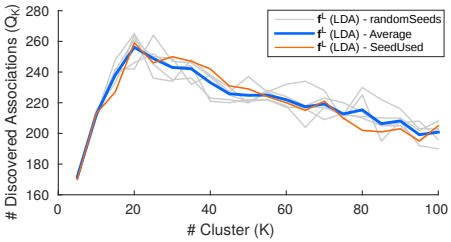


**Fig. 3.** Number of discovered associations ($Q$) over varying numbers of clusters ($K$) for different seeds (gray) and averaged (blue). Red indicates the seed used to generate the evaluation results.

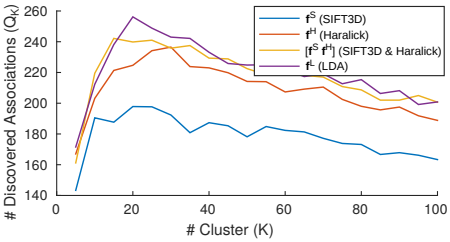


**Fig. 4.** Comparison of different feature sets. "SIFT3D+Haralick" denotes the concatenation of the two feature sets. The values represent the average using 5 different seeds.

number of clusters $K$ we define the measure of quality

$$Q_K = \sum_{k=1}^{K} \sum_{t=1}^{T} [p_{tk} \leq 0.05]. \tag{2}$$

## 4 Results

Fig. 3 shows values of the quality criterion (Eq. 2) for various numbers of $K$ using the LDA volume descriptor $\mathbf{f}^L$ for clustering. K-means is based on random initialization. Thus, to rule out random effects, we perform the experiments with a set of 5 different random seeds. Graphs are shown for each seed (gray), the average result (blue) that was used to determine the number of clusters and the random seed (red) for which the evaluation results are reported. Fig. 4 shows a comparison of different feature sets ($\mathbf{f}^H$, $\mathbf{f}^S$ and $\mathbf{f}^L$) with respect to the clustering quality $Q_K$. Concatenating texture and shape features $[\mathbf{f}^H\ \mathbf{f}^S]$ allows to discover more structure in the data than each feature set individually. The LDA embedding $\mathbf{f}^L$ further improves the number of associations discovered. For further results the descriptor $\mathbf{f}^L$ and the number of cluster 20 are fixed. Fig. 5 illustrates the visual variability of the data by showing a 2D visualization of the $\mathbf{f}^L$ descriptors using t-SNE [12]. In addition, exemplary slices of volumes at different positions in the feature space are shown. Fig. 6a illustrates all associations discovered by population clustering. Positive associations ($OR_{tk} > 1$) and negative associations ($OR_{tk} < 1$) are shown for all $p_{tk} \leq 0.05$. Fig. 6(b-e) shows a comparison of 3 exemplary clusters illustrating the raw features (b), the embedding (c) a set of terms that are associated with the cluster (d) and exemplary slices of volumes in the cluster (e).

## 5 Conclusion

We propose a framework for visual population clustering of large clinical routine imaging data. After spatial normalization, visual features are learned, and a

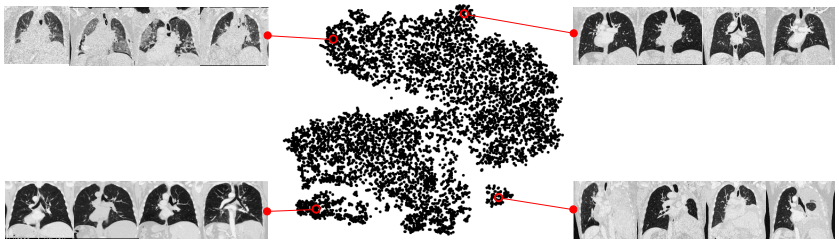

**Fig. 5.** 2D visualization of the LDA image descriptors of 7812 volumes using t-SNE. Exemplary volume slices from different areas in the feature space are given to illustrate the visual variability in the population.

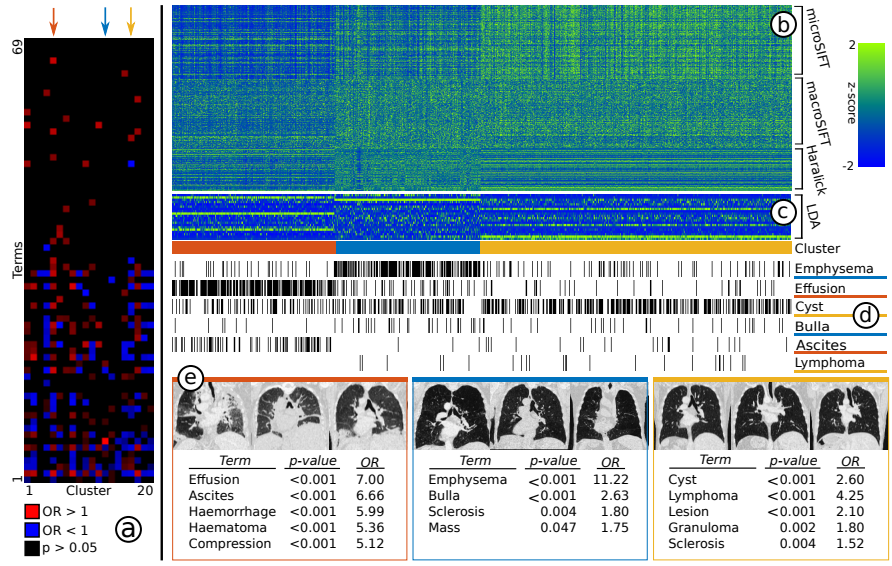


**Fig. 6.** (a) Discovered associations between clusters (columns) and terms (rows). Terms are sorted by decreasing occurrence frequency. Positive associations ($OR > 1$) are indicated red and negative associations ($OR < 1$) are indicated blue. (b-e) comparison of three clusters. (b) shows raw features, (c) the LDA embedding and (d) indicates the appearance of 6 terms that are overrepresented in one of these clusters. (e) shows exemplary volume slices of members and lists of up to 5 significantly overrepresented terms with p-values and OR of the respective clusters.

clustering is performed on the volume level. We evaluate the impact of features on the clustering, and validate the clinical relevance of the resulting grouping of patients based on corresponding radiology reports. Results show that the clustering after normalization identifies groups with coherent sets of reported findings. This demonstrates that visual markers that relate to clinical findings, can be learned without supervision. The proposed approach is a step towards unsupervised learning from clinical routine imaging data.

## References


1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature communications 5 (2014)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of machine Learning research 3, 993–1022 (2003)
3. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: Images are more than pictures, they are data. Radiology 278(2), 563–577 (2016)
4. Göksel, O., Jiménez-del Toro, O.A., Foncubierta-Rodríguez, A., Muller, H.: Overview of the VISCERAL challenge at ISBI. In: Proceedings of the VISCERAL Challenge at ISBI. New York, NY (2015)
5. Gruslys, A., Acosta-Cabronero, J., Nestor, P.J., et al.: A new fast accurate nonlinear medical image registration program including surface preserving regularization. IEEE Transactions on Medical Imaging 33(11), 2118–2127 (2014)
6. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics 3(6), 610–621 (1973)
7. Hofmanninger, J., Langs, G.: Mapping visual features to semantic profiles for retrieval in medical imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 457–465 (2015)
8. III, W.M.W.: A feature-based approach to big data analysis of medical images. In: Proceedings of the 24th Int. Conf. on Information Processing in Medical Imaging, IPMI. vol. 9123, p. 339. Springer (2015)
9. Kumar, R.S., Senthilmurugan, M.: Content-based image retrieval system in medical applications. Int. Journal of Engineering Research and Technology 2(3) (2013)
10. Kumar, V., Gu, Y., et al.: Radiomics: the process and the challenges. Magnetic resonance imaging 30(9), 1234–1248 (2012)
11. Lambin, P., Rios-Velazquez, E., Leijenaar, R., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European Journal of Cancer 48(4), 441–446 (2012)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(2579-2605), 85 (2008)
13. Mondal, P., Mukhopadhyay, J., Sural, S., Bhattacharyya, P.P.: 3d-sift feature based brain atlas generation: An application to early diagnosis of alzheimer's disease. In: Int. Conf. on Medical Imaging, m-Health and Emerging Communication Systems. pp. 342–347. IEEE (2014)
14. Shin, H.C., Lu, L., Kim, L., Seff, A., Yao, J., Summers, R.M.: Interleaved text/image deep mining on a very large-scale radiology database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1090–1099 (2015)
15. Toews, M., Wells, W.M.: Efficient and robust model-to-image alignment using 3d scale-invariant features. Medical image analysis 17(3), 271–282 (2013)
16. Vogl, W.D., Prosch, H., Müller-Mang, C., Schmidt-Erfurth, U., Langs, G.: Longitudinal Alignment of Disease Progression in Fibrosing Interstitial Lung Disease. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). vol. 8671, pp. 97–104. Springer International Publishing (2014)
17. Wibmer, A., et al.: Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. European radiology 25(10), 2840–2850 (2015)