Explorative Hyperbolic-Tree-Based Clustering Tool for Unsupervised Knowledge Discovery

Michael Riegler[1], Konstantin Pogorelov[1], Mathias Lux[2], Pål Halvorsen[1], Carsten Griwodz[1]
Thomas de Lange[3], Sigrun Losada Eskeland[4]
[1]Simula Research Laboratory, Norway
[2]Klagenfurt University, Austria
[3]Vestre Viken Hospital Trust, Bærum Hospital, Cancer Registry of Norway, Norway
[4]Department of Medical Research, Bærum Hospital, Vestre Viken Health Trust, Norway

***Abstract*—Exploring and annotating collections of images without meta-data is a laborious task. Visual analytics and information visualization can help users by providing interfaces for exploration and annotation. In this paper, we show a prototype application that allows users from the medical domain to use feature-based clustering to perform explorative browsing and annotation in an unsupervised manner. For this, we utilize global image feature extraction, different unsupervised clustering algorithms and hyperbolic tree representation. First, the prototype application extracts features from images or video frames, and then, one or multiple features at the same time can be used to perform clustering. The clusters are presented to the users as a hyperbolic tree for visual analysis and annotation.**

I. Introduction

Content-based image retrieval has been an important area of research for quite some time now [1]. A lot of different techniques and methods have been created, and the approaches have become more and more sophisticated. However, there is no one-fits-all approach, and the tools often must be adapted to a particular use-case.

One of the domains we are focusing on is medical images from the human gastrointestinal tract, taken with an endoscope camera inside the body to detect diseases. Even though these images are coming from a particular patient and have been annotated by a particular endoscopist, the domain is not as meta-data rich as intuitively anticipated. Highly trained and specialized medical personnel are scarce human resources, and their priority is on performing medical examinations, not annotating or giving sense to images and videos [2], [3]. Moreover, if videos and frames are shared, the patients personalized information has to be purged from this data or anonymized to ensure privacy of the patients, and especially, in case of shared videos and frames from endoscopic procedures, meta-data is a rare commodity. Therefore, a lot of videos and video frames remain only loosely annotated, and retrieving the images later based on available information is hard.

In this context, we present a prototype mainly designed for visual analysis and annotation of endoscopic images. The prototype application has two main benefits. First, it allows clinical personnel to investigate and analyze vast collections of frames from endoscopic procedures by providing a configurable focus and context view based on frame similarity. Second, it allows for utilizing the focus and context view for annotation and tagging of the dataset, making it more accessible for complementary information systems. While we developed this prototype application for a medical scenario, we strongly believe, and will also show in the evaluation, that it is usable for other scenarios involving interactive browsing, visual analysis or annotation of image or video data. We first investigate the relation between focus and context views and content-based image similarity, as well as discuss the underlying frameworks of the application. We then pick two diverse datasets, one from the medical domain and one from social image collections, to investigate if the proposed abstraction and clustering of the images is applicable through an evaluation. Then, we describe our prototype and show how it can be used to support professional users in the domain of analysis of endoscopic video frames in their daily work routine. Finally, we discuss the contribution of the application and further work on the topic.

II. Related work

Chi [4] defines information visualization in four stages (Table I). First, *raw data* is transformed into an *analytical abstraction*, which is transformed into a *visualization abstraction*, which itself then is presented in a *view*. As indicated in Table I, the data we operate on is images, and for the view stage, we chose a hyperbolic tree visualization.

TABLE I
Prototype stages of visualization and correspondence.

	Stage	In our prototype
1	Raw data	Images/ Video frames
2	Analytical abstraction	Image feature descriptors
3	Visualization abstraction	Clusters, centroids and distance values
4	View	Hyperbolic tree

One of the first and most prominent of these approaches was the hyperbolic browser by Lamping, Rao and Pirolli [5]. The underlying idea is, that the visualization abstraction is based on a hierarchy, i.e., a directed tree. In a typical view, the objects would be arranged in a certainly, with those in focus being larger and closer to the center, while those not in

978-1-4673-8695-1/16/$31.00

focus, i.e., the ones being the context, are pushed to the rim of the circle. A hyperbolic view on a hierarchical structure is best described with a fish eye view on a particular tree branch or leaf, with the rest being visible, but out of focus.

The hyperbolic tree visualization is a graph based information visualization strategy [6], which has been applied mostly to data that already closely resembles a tree structure or a directed graph from which a tree can be abstracted including hypertext collections like the WWW, social networks, ontologies and other data where transformation between raw data and abstraction remains on a low complexity level. One of the few examples, where image collections are interpreted as graph structure based on their content, is presented in [7], where the authors employ a force directed placement algorithm to display images on a large video wall. Without the focus and context view, however, the authors are limited by the size of the video wall. Other work of the same authors focuses on displaying images based on content based similarity in a Treemap [8]. The *PhotoTOC* project [9], on the other hand, used clustering to create an *overview+detail view* by clustering images based on color histograms and then presenting the clusters by their medoids. In [10], images are displayed based on their distance with respect to two shape and texture features. Clustering does not take place, but the focus of the visualization lies on the query image and the k nearest neighbors. The rest of the result list is pushed to the outer rim of the visualization providing a context.

III. Analytical and Visual Abstraction

The features for clustering, i.e., the analytical abstraction as defined in Table I, are extracted with LIRE (latest modified version[1]). LIRE supports multiple global and local features out of the box, to allow for easy integration of features in arbitrary applications. Most notable global ones are the Color and Edge Directivity Descriptor (CEDD) [11] as well as the related features including the Joint Composite Descriptor (JCD) [12], the Fuzzy Color and Texture Histogram (FCTH) [13], the Pyramid Histogram of Oriented Gradients (PHOG) [14], the Auto Color Correlogram [15], Local Binary Patterns [16], CENTRIST [17]. Additionally, it includes the MPEG-7 features [18] Edge Histogram, Color Layout and Scalable Color. A detailed description of the extraction process and the features can be found in [19].

For the visualization abstraction stage (see table I), we use WEKA [20]. WEKA is a collection of tools for machine learning and data mining providing also a Java library, which can be directly combined with the LIRE code for our prototype. In the fusion between these two frameworks, LIRE is responsible for the feature extraction and also for the main program logic calling the required functions from WEKA. The coupling allows for optional change of the employed clustering routine. For the experiment described in this paper, the *X-means* clustering algorithm [21] is used, because X-means determines the number of the clusters automatically, which is an important part of the experiment. Our demo also supports K-means and hierarchical clustering [22].

One of the main aspects of our demo is interactivity with the view, i.e., users interact with the created clusters. Clustering, being a well-known technique in machine learning, is used to group entities based on a similarity metric. For instance, images can be group-based on image features (e.g., grouping those with similar colors), or textual user comments can be clustered based on the nouns they contain. For our demo, we use two datasets. One to group pictures showing disease symptoms in a medical scenario, the other to group pictures of the same tagging categories in a social image collection. With visual analysis, these clusters can be investigated by users with domain knowledge about the images content to confirm or reject the grouping within an annotation process.

While being developed for a medical scenario, our prototype is not restricted to a specific domain. Taking advantage of this, we first investigate the appropriateness of the analytical abstraction stage, i.e., the selection of features, as well as the visualization abstraction stage, i.e., the clustering, using two very different publicly available datasets. The first one is the intent dataset of Lux et al. [23]. This dataset contains $1,310$ images crawled from Flickr as well as results from a survey regarding the intentions of the photographers and responses from the photographers as well as crowd-workers judging the images and annotations. The intent categories, from which the users had to choose, are (i) *preserve a good feeling*, (ii) *preserve a bad feeling*, (iii) *show it to family and friends*, (iv) *publish it on-line*, (v) *support a task of mine* and (vi) *recall a specific situation*. For this dataset, the experiment is done for single global features as well as for feature fusions. The second dataset is the ASU-Mayo Clinic polyp dataset which is the biggest publicly available dataset for polyp detection in medical images consisting of 20 videos, with a total number of $18,781$ image frames [24].

On both datasets, we conducted two-step experiments which are slightly different in their final evaluation metric. The first step is clustering the images with our tool based on their global features. The number of clusters is not predetermined, but suggested by X-means. This step is identical for both datasets. For the intent dataset, the mean squared error is then calculated per cluster. In our evaluation, the correlation between the users' feedback and the mean square error of the clusters is computed for the intent dataset. If the correlation coefficient ρ is low, i.e., close to -1, we assume that the method works well, as inter-user-agreement is high while mean square error is low, or the other way around. ρ around 0 or a positive ρ near 1 would indicate that mean square error and user agreement are either not correlated or correlated in the wrong way, implying that the clustering does not work. The intent dataset contains votes of three different users for each category. The users indicates on a 5-point Likert scale how representative an image is for a given category (1, strongly disagree, to 5, strongly agree). For all user votes, the majority vote is calculated and all of them are averaged and normalized.

For the ASU dataset, we can not calculate the mean squared

[1]https://github.com/dermotte/lire, last visited 2016-03-08

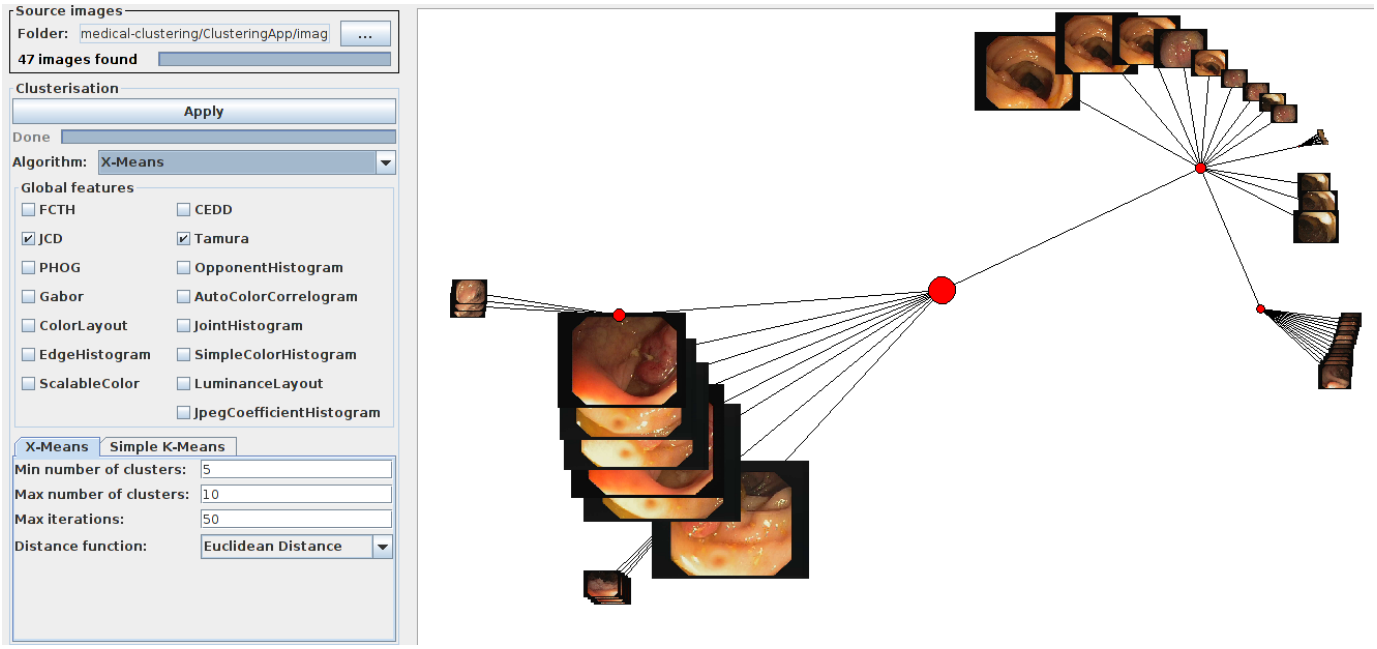

Fig. 1. Demo system: The *left* part contains the settings for the users, and the *right* part shows the output of the clustering as a hyperbolic tree.

error because it contains only binary classification for each frame: a polyp is visible in the image or not. Instead, we calculated the purity of the clusters based on the ground truth provided with the dataset. Furthermore, while we used single global features for the intent dataset, which have been report to work well, we used a combination of the JCD and Tamura features for the ASU dataset. These have been found to work best for this dataset based on an information gain analysis.

Table II shows the results of the experiment based on the intent dataset. As expected, a negative correlation is observed, which means that the clustering results correlate with manual annotations to a degree indicated by the absolute value of ρ. At first, it shows that some global features are more suitable to create clusters that are similar with user judgments than others. For example, FCTH is the best feature for detecting a *publish on-line* intent for an image. A closer look at the clusters generated by FCTH shows that this feature can very well detect if persons are shown in an image, and it seems that most images used for on-line publishing contain one or more persons. Another interesting insight is that semantically similar clusters are also correlated similar to the same feature, e.g., Gabor features for *recall situation* and *preserve good feeling*. This is also an indication that a combination of features is more suitable to provide clusters that are consistent with with user judgments. The last important insight, which is given by this first experiment, is that a simple combination of all features does not automatically lead to better correlation. This indicates that the right choice of feature combinations is important for clustering and that a metric like information gain can give an idea about what features to combine, which we also used in our next experiment. The second experiment with the ASU dataset revealed something similar to the previous experiment. First, we performed information gain analysis to identify the two best features for this dataset. This led us to the features JCD and Tamura, which we combined using early fusion. Based on these features, we performed 4 different tests with different numbers of clusters. We used X-means to determine four clusterings of the dataset. We let X-means determine the number of clusters c for one experiment, then we clustered with $c \in \{2, 4, 100\}$. Based on the created clusters, we calculated the average purity (precision based on the majority class for each cluster). For c equals 2, 4 and 100, we got a purity of 77%, 97% and 95%, respectively. For $c = 234$, the c proposed by the X-means algorithm, the purity is 97%. This indicates that the clustering leads to meaningful results also for the ASU dataset and therefore supports our approach for analytical and visualization abstraction.

TABLE II
CORRELATION ρ BETWEEN MEAN SQUARED ERROR AND USER VOTES FOR DIFFERENT GLOBAL FEATURES OF THE INTENT DATASET [23].

Feature	recall	preserve good	publish	show	support	preserve bad
CEDD	0,165	0,194	0,205	0,285	0,213	-0,05
FCTH	0,085	-0,11	-0,70	-0,32	0,298	-0,27
Gabor	-0,50	-0,40	-0,03	-0,15	-0,08	0,254
Tamura	-0,77	-0,24	0,050	-0,55	0,241	0,517
Luminance Layout	0,060	-0,32	-0,15	-0,30	0,002	0,248
Scaleable Color	0,126	0,295	-0,02	0,060	-0,05	0,094
Opponent Histogram	0,107	-0,07	-0,10	-0,03	0,085	-0,003
AutoColor Correlogram	0,691	0,609	0,739	0,779	-0,47	-0,67
JPEG Coefficent	-0,10	0,006	-0,26	-0,04	-0,48	0,107
Edge Histogram	-0,17	0,643	-0,26	-0,06	-0,51	-0,04
PHOG	-0,52	0,225	0,024	-0,42	0,187	-0,06
JCD	0,168	0,288	0,227	0,193	0,275	-0,26
JointHistogram	0,408	0,262	0,447	0,238	0,396	-0,40
12 Features Combined	-0,14	0,469	-0,11	-0,17	0,215	0,735

IV. PROTOTYPE AND DEMO

Our prototype application combines content-based similarity, unsupervised classification and focus/context views to provide a way to easily explore, analyze and annotate a vast number of video frames or images. Figure 1 shows a screen

shot of the demo application. On the upper left side, users can choose the folder containing the image collection. Below that, the clustering algorithm can be selected. At the moment, we support 3 different algorithms (K-means, X-means and hierarchical clustering). After selecting the clustering algorithm, the application allows to choose one or several different image features. For the screen shot, we limited the list, but the final demo will contain all of the image features provided by LIRE. If more than one feature is picked, they will be combined using early fusion. The final options allow the user to specify the clustering parameters. As a default, we use the values recommended by WEKA. After the users choose the images and all the options, a click on **Apply** creates the clusters and presents them as a hyperbolic tree on the right site. The cluster leaves are represented using the image that is closest to the cluster center, i.e., the cluster medoid. It is possible to interact with the tree by zooming and turning it into different angles. Furthermore, the user can double click on images, which will open the folder containing all images in the selected cluster. A right click on the cluster images allows the user to see information like the cluster center and the purity of the cluster based on the distances. Finally, the users can name/tag the clusters, which adds the tag to the name of the images in the cluster (in this format _"your tag".filetype). For the demo, we will present how our tool works on the two different datasets that we tested here, but we will also have a new large dataset of different endoscopic findings that we will use during the demo presentation.

V. Conclusion

In this paper, we presented a demo application that enables domain experts to use unsupervised clustering algorithms to explore image and video data collections that do not contain meta-data. In the information visualization model of the four stages, the analytical abstraction stage and the visualization abstraction stage correspond to the selection and extraction of image features and the clustering of the feature vectors. We have shown – based on two different datasets – that the clustering leads to good results which correspond to user judgments or ground truth of the datasets, and therefore, provide good candidate methods for the abstraction stages.

For future work, we plan to test the application with domain experts. In our case, endoscopists from two different Norwegian Hospitals. For this test, we already collected a large dataset (200.000 images and 600 videos) from medical procedures. Focus of this user study will be the usefulness of the focus+context view as well as the perceived complexity of the user interface, i.e., the selection of image features and clustering algorithms.

Acknowledgment

This work is funded by the "EONS" FRINATEK project (231687).

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] M. Riegler, K. Pogorelov, P. Halvorsen, T. de Lange, C. Griwodz, P. T. Schmidt, S. L. Eskeland, and D. Johansen, "EIR - efficient computer aided diagnosis framework for gastrointestinal endoscopies," in *Proc. of CBMI*, 2016.

[3] K. Pogorelov, M. Riegler, P. Halvorsen, P. T. Schmidt, C. Griwodz, D. Johansen, S. L. Eskeland, and T. de Lange, "GPU-accelerated real-time gastrointestinal diseases detection," in *Proc. of CBMS*. IEEE, 2016.

[4] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in *Proc. of IEEE InfoVis*, 2000, pp. 69–75.

[5] J. Lamping, R. Rao, and P. Pirolli, "A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies," in *Proc. of SIGCHI conf. on Human factors in comp. sys.*, 1995, pp. 401–408.

[6] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Trans. on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[7] Y. Gu, C. Wang, J. Ma, R. J. Nemiroff, and D. L. Kao, "igraph: a graph-based technique for visual analytics of image and text collections," in *IS&T/SPIE Electronic Imaging*, 2015, pp. 939 708–939 708.

[8] C. Wang, J. P. Reese, H. Zhang, J. Tao, and R. J. Nemiroff, "imap: A stable layout for navigating large image collections with embedded search," in *IS&T/SPIE Electronic Imaging*, 2013, pp. 86 540K–86 540K.

[9] J. C. Platt, M. Czerwinski, and B. A. Field, "Phototoc: Automatic clustering for browsing personal photographs," in *Proc. of ICICS-PAM*, 2003, pp. 6–10.

[10] R. S. Torres, C. G. Silva, C. B. Medeiros, and H. V. Rocha, "Visual structures for image browsing," in *Proc. of ACM CIKM*, 2003, pp. 49–55.

[11] S. A. Chatzichristofis and Y. S. Boutalis, "Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval," in *Computer Vision Systems*. Springer, 2008, pp. 312–322.

[12] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux, "Selection of the proper compact composite descriptor for improving content based image retrieval," in *Proc. of IASTED SPPRA*, 2009.

[13] S. Chatzichristofis, Y. S. Boutalis *et al.*, "FCTH: Fuzzy color and texture histogram-a low level feature for accurate image retrieval," in *Proc. of IEEE WIAMIS*, 2008, pp. 191–196.

[14] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. of ACM CIVR*, 2007, pp. 401–408.

[15] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. of IEEE CVPR*, 1997, pp. 762–768.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[17] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[18] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the mpeg-7 standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001.

[19] M. Lux and G. Macstravic, "The LIRE request handler: A Solr plug-in for large scale content based image retrieval," in *Proc. of MMM*, Dublin, IE, Jan 2014, pp. 374–377.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *ICML*, vol. 1, 2000, pp. 727–734.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[23] M. Lux, M. Taschwer, and O. Marques, "A closer look at photographers' intentions: a test dataset," in *Proc. of ACM MM workshops - Crowdsourcing for multimedia*, 2012, pp. 17–18.

[24] N. Tajbakhsh, S. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.