Visual Phrase Learning and Its Application in Computed Tomographic Colonography

Shijun Wang, Matthew McKenna, Zhuoshi Wei, Jiamin Liu, Peter Liu, and Ronald M. Summers

Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bldg 10, Room 1C224, Bethesda, MD 20892-1182, U.S.
rms@nih.gov

Abstract. In this work, we propose a visual phrase learning scheme to learn an optimal visual composite of anatomical components/parts from CT colonography images for computer-aided detection. The key idea is to utilize the anatomical parts of human body from medical images and associate them with biological targets of interest (organs, cancers, lesions, etc.) for joint detection and recognition. These anatomical parts of the human body are not necessarily near each other regarding their physical locations, and they serve more like a human body navigation system for detection and recognition. To show the effectiveness of the proposed learning scheme, we applied it to two sub-problems in computed tomographic colonography: teniae detection and classification of colorectal polyp candidates. Experimental results showed its efficacy.

1 Introduction

To help radiologists read images and identify lesions, various computer-aided detection (CADe) and computer-aided diagnosis (CADx) systems have been developed [1, 2]. In the majority of CAD systems developed for radiology, anatomical knowledge is highly embedded into the algorithm design. In other words, interaction with radiologists during algorithm development, and integration of expert human knowledge into the algorithm, is crucial to the success of these CAD systems. However, instead of relying on a radiologist to define the anatomical knowledge used in a CAD system, we believe that a computer could learn what parts of the human anatomy are useful in performing the detection task. Particularly, this work focuses on how to automatically build an anatomical model of human body using only statistical information from CT images.

In recent years, significant progress has been made in the field of computerized object detection and recognition due to the application of statistical learning on large-scale data. Some cutting edge methods include bag of words (BoW) [3], deformable templates [4], and part-based models [5]. BoW methods are built from codebooks, or collections, of visual patches extracted from images. BoW methods usually employ affine invariant descriptors to characterize image patches. Furthermore, the efficient

K. Mori et al. (Eds.): MICCAI 2013, Part I, LNCS 8149, pp. 243–250, 2013.

creation of useful codebooks for visual object recognition is a critical step in BoW methods [5]. In the deformable templates technique, the key idea is to fit a model to an image by minimizing the error between the input image and the closest model instance. Finally, with part-based models, the target object is modeled by mixtures of multi-scale deformable part models [5]. In the work of [6] on part-based models, they proposed an explicit way to utilize auxiliary/accompanied objects to detect and recognize a main target object. They called their approach visual phrase recognition [6], where a visual phrase is a complex visual composite containing several objects (i.e. "a person riding a horse").

In this work, we propose a visual phrase learning scheme to learn visual composites in medical images. The key motivation is to develop an automatic way to learn visual phrases, instead of the manual way in the original work [6]. We utilize the anatomical parts of human body from medical images by composing them with a target of interest (e.g. organs, cancers, lesions) for joint detection and recognition. Unlike the work of [6], the relevant anatomical parts of human body are not necessarily near each other, and they serve more as a human body navigation system for detection and recognition. To show the efficacy of our new visual phrase, we applied our learning scheme to two problems encountered in CTC: the identification of the teniae coli and the classification of polyp candidates in a CAD system (Fig.1).

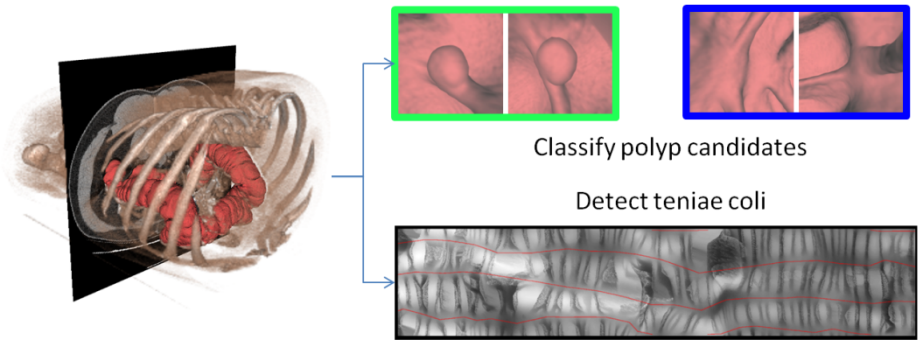

Fig. 1. Overview of CTC and two applications: colonic polyp classification (green square shows a true polyp and blue square shows a false positive) and teniae coli detection

2 Visual Phrase Learning

We show a diagram of our proposed system in Fig. 2. In the work of Sadeghi and Farhadi [6], the visual phrases were determined by prior knowledge using bounding boxes. All components belonging to a visual phrase were in close spatial proximity, making the system a top-down approach in which a visual phrase is determined beforehand and then applied to test images. What happens if we do not know the visual parts of the phrase a priori, or if we have a large pool of visual parts (codebook) and do not know which combination will be helpful for the detection of the target object? Also, how can visual phrases be developed for visual parts that are distributed in different spatial locations? In the following subsections, we address these problems

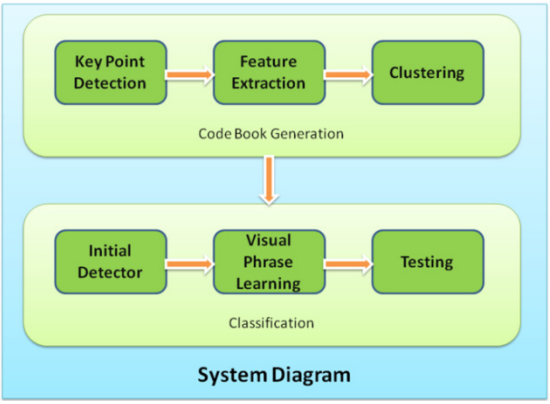

Fig. 2. System diagram of the proposed visual phrase learning algorithm

by proposing a bottom-up method to learn the optimal visual phrase for joint object detection and recognition from a codebook of visual parts.

2.1 Problem Formulation

For a two-class classification problem, given training samples $\{(X_1,y_1),\ldots,(X_n,y_n)\}$, $y_i \in \{-1,+1\}$, where each training sample X is a composition of detection x and auxiliary data (visual patches coming from a codebook with m items) associated with each detection $\{x_1^c,\ldots,x_m^c\}$, the visual phrase learning problem is formulated as follows:

$$\min_{w_c} \min_{w_1,b,\xi} \frac{1}{2}\|w_1\|^2 + C_1 \sum_{i=1}^{n} \xi_i \,, \tag{1}$$

$$\text{s.t. } y_i\left(w_1^T X_i w_c + b\right) \geq 1-\xi_i, \xi_i \geq 0, \forall_{i=1}^{n}, w_c \geq 0, \sum_{i=0}^{m} w_{ci} = 1,$$

where w_1 is a linear classifier in feature space; b is the bias item of the classifier; $w_c \geq 0$ is a parameter vector to learn visual phrase from a codebook; and C_1 is a trade-off parameter to control the classifier complexity and slack variables ξ_i, $i = 1,2,\ldots,n$; Please note that each training sample, X_i, is a $d \times (m + 1)$ matrix in which each column represents a detection/instance, each row corresponds to a feature and d is the feature's dimension. The first column is the detection we wish to classify. The next m columns represent m parts from a code book. The column-wise order of detection and code book items is fixed to maintain consistency across all training and test samples. The purpose of the above learning problem is to learn the visual phrase (specified by w_c) which has the best performance regarding classification, given a codebook from the image data. Our hypothesis is that a visual phrase (composed of detection and its surrounding structures) has better discriminating power than a single detection object. Eq. (1) is a new formulation which contains the key idea on visual phrase learning proposed in this paper. Please note that X_i in Eq. (1) is a 2D matrix which differentiates the proposed formulation from traditional support vector machines formulation.

2.2 Linear SDP Solution

To solve the above optimization problem, we propose the following theorem:

Theorem 1. The above visual phrase learning problem can be formulated as the following semi-definite programming (SDP) problem:

$$\min_{w_c,\ \nu,\delta,\lambda} t \qquad \text{s.t.} \begin{bmatrix} K & \frac{(e+\nu-\delta+\lambda y)}{\sqrt{2}} \\ \frac{(e+\nu-\delta+\lambda y)^T}{\sqrt{2}} & (t-C_1\delta^T e) \end{bmatrix} \geq 0, \tag{2}$$

$$W_c = W_c^T,\ e^T W_c e = 1,\ \nu \geq 0, \delta \geq 0,$$

where $K_{ij} = y_i y_j trace\left(X_i^T X_j \times W_c^T\right); \nu \geq 0, \delta \geq 0$ and λ are dual variables introduced in the dual problem; e is a vector filled with all ones. Our SDP solution provides a closed-form solution to the optimization problem in (1), which is guaranteed to be global optimal. The proof is omitted due to page limit.

It is interesting to note that the solution we show in Theorem 1 has connections with multi-kernel learning [7] and sequence kernel learning [8]. Multi-kernel learning can be viewed as special cases of our visual phrase learning framework.

2.3 Kernelization

In the previous subsection we showed the linear solution for the visual phrase learning problem in the original feature space. Now let's consider its nonlinear solution. First let us define a mapping function Φ which maps the data in the original Euclidean space to a new reproducing kernel Hilbert space (RKHS): $\mathbb{R}^d \rightarrow H$. More specifically, the mapping function Φ maps each column of the input sample (detection plus codebook items) to the same RKHS. H may be infinite dimensional. Utilizing a different mapping function and RKHS for detection and codebook is also feasible but beyond the scope of this paper. The visual phrase learning problem in the new Hilbert space can be formulated as follows:

$$\min_{w_c} \min_{w_1,b,\xi} \frac{1}{2}\|w_1\|^2 + C_1 \sum_{i=1}^{n} \xi_i, \tag{3}$$

$$\text{s.t. } y_i\left(w_1^T \Phi(X_i) w_c + b\right) \geq 1-\xi_i,\ \xi_i \geq 0,\ \forall_{i=1}^{n}, w_c \geq 0, \sum_{i=0}^{m} w_{ci} = 1.$$

$\Phi(X_i)$ is the mapping of detection x_i and its corresponding codebook items. Each column of $\Phi(X_i)$ corresponds to one vector in the new RKHS. We define a symmetric kernel function for the input samples (detection plus codebook items) as follows: $K_f\left(X_i, X_j\right) = trace\left(\Phi(X_i)^T \Phi\left(X_j\right) \times W_c^T\right)$. Let us define the corresponding symmetric kernel function for the mapping function Φ as follows: $K_c\left(X_{im}, X_{jn}\right) = \Phi(X_{im})^T \Phi\left(X_{jn}\right)$

where subscripts m and n correspond to m'th and n'th columns of sample X_i and X_j. Then we have the following Theorem:

Theorem 2. The symmetric kernel function $K_f(X_i, X_j) = trace\left(\Phi(X_i)^T \Phi(X_j) \times W_c^T\right)$ fulfills the Mercer's condition for any kernel function $K_c(X_{im}, X_{jn}) = \Phi(X_{im})^T \Phi(X_{jn})$ fulfilling the Mercer's condition when $w_c \geq 0$ (element-wise).

Proof: For any square integrable functions g(x),

$$\iint K_f(X_i, X_j) g(X_i) g(X_j) dX_i dX_j = \iint trace\left(\Phi(X_i)^T \Phi(X_j) \times W_c^T\right) g(X_i) g(X_j) dX_i dX_j$$

$$= \sum_{u=0}^{m} \sum_{v=0}^{m} w_{cu} w_{cv} \iint \Phi(X_{iu})^T \Phi(X_{jv}) g(X_i) g(X_j) dX_i dX_j \geq 0 \qquad \square$$

By using the Lagrange multiplier optimization method, we obtain the theorem:

Theorem 3. The nonlinear visual phrase learning problem can be formulated as the following semi-definite programming (SDP) problem:

$$\min_{w_c,\ \nu, \delta, \lambda} t \qquad \text{s.t.} \begin{bmatrix} K & \dfrac{(e + \nu - \delta + \lambda y)}{\sqrt{2}} \\ \dfrac{(e + \nu - \delta + \lambda y)^T}{\sqrt{2}} & (t - C_1 \delta^T e) \end{bmatrix} \geq 0, \qquad (4)$$

$W_c = W_c^T$ and $e^T W_c e = 1$, $\nu \geq 0, \delta \geq 0$, where

$K_{ij} = y_i y_j trace\left(\Phi(X_i)^T \Phi(X_j) \times W_c^T\right)$. Proof is omitted due to the page limit.

3 Experiments: Teniae Coli

Teniae coli are three longitudinal smooth muscle bands in the colon surface. They are parallel, equally distributed, and form a triple helix structure from the appendix to the sigmoid colon. Fig.1 illustrates a human colon and the configuration of the teniae coli. Teniae are anatomically meaningful landmarks and can be used to estimate the circumferential positions of potential lesions in CT colonography.

To detect teniae coli, colon segmentation was first performed on the original CTC slice. The segmented colon surface was reconstructed and unfolded into a 2D flattened colon using a reversible projection. The unfolded images then were converted into 2D height maps. The height maps are 2D intensity images that record the elevation of the colon surface relative to the unfolding plane, where haustral folds correspond to high elevation points and teniae to low elevation points.

We used CTC data from 20 patients, dividing patients into separate training (17) and testing (3) sets. The separation of training/test sets was determined empirically. We cropped each image to only include the middle segment of each image as the teniae were very difficult to define at either end of the colon.

To generate the codebook used in our approach, we utilized a multiscale Harris operator to detect points of interest on our images [9]. The detector located corners and edges at various scales. We extracted a small patch around each detected point. Fig. 3(a) shows some typical keypoints detected by the Harris operator. Note that these visual words are not adjacent to each other. We utilized a range of features to describe each keypoint, including HOG, shape context, and SIFT. We also included the location of each detection and a histogram of intensity values as additional features. A k-means clustering scheme (k=5) was used to learn the elements in our codebook. The number of visual words was empirically determined based on performance of the system on the training set. For kernel computation and classification, we used a radial basis function (RBF) kernel with a kernel width parameter set as the 90th percentile of pairwise distances between all training samples. We applied the same kernel to the following colonic polyp detection data. The RBF kernel was chosen because its efficacy on many real applications.

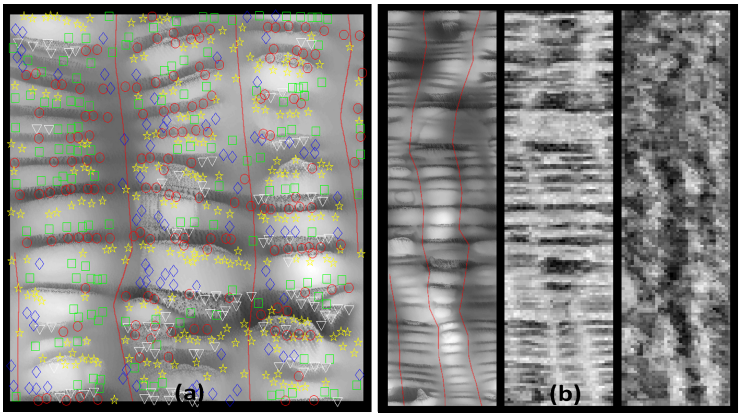

Fig. 3. (a) Keypoints from subset of a flattened colon from a CTC scan, detected using a Harris detection scheme. Different markers denote different clusters. Note the regular pattern around the folds. (b) Detection results of teniae coli. From left to right: original height map of a colon with ground truth (red lines); detection results of using detection features only; detection results using the proposed visual phrase approach. The brighter the detection block, the higher the probability it is a tenia coli.

In Fig. 3(b) we show comparisons between the proposed method and an SVM that used detection features only. As is evident in the above figure, the visual phrase classifier is able to learn the pattern of the teniae falling between and orthogonal to folds. All the non-teniae structures, like folds and colon walls, were suppressed by the visual phrase. The SVM, which is blind to surrounding structures, mistakes areas between folds (running horizontally, orthogonal to the teniae) as teniae.

4 Experiments: Colonic Polyp Detection in CTC

In the polyp detection problem, we used CTC data from 50 patients, dividing patients into separate training (25) and testing (25) sets. We extracted the colon from each

scan and performed a curvature-based analysis to generate an initial list of lesions of interest. After initial filtering, we identified 880 polyp candidates (detections) which include 62 true polyps.

For each polyp candidate, our system generated 5 intraluminal, volume-rendered images focusing on the detection from various viewpoints (Fig. 1 shows two viewpoints as illustration). Averaged prediction scores of the 5 images were used as the final prediction score for the polyp candidate. We used 2D HOG features to describe the images (the first column of sample X in Section 2.1). To generate a codebook of auxiliary data, traditional 3D curvature based features were extracted from original CT slices. These 3D features are widely used in traditional CTC CAD to capture the anatomical shapes of the polyp candidate and its surrounding structures.

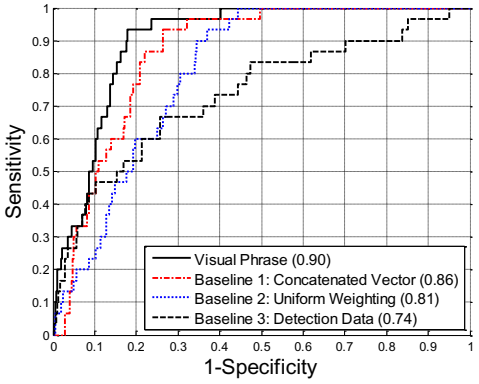

Fig. 4. ROC comparison of the proposed visual phrase method and baseline methods. The difference between the visual phrase method and the highest-performing baseline method (concatenated vector) is significant (p<0.05).

We produced receiver operating curves (ROC) to analyze the performance of the visual phrase. We compared the visual phrase with 3 baseline methods. For the first method, we concatenated the detection data and all codebook items into a single vector that was fed to the classifier. For the second method, we set the Wc matrix to equally weight the detection data and each codebook item. We also compared the visual phrase classifier to a system that only used the detection features. In Fig. 5 the ROC's of the four methods are compared. Use of visual phrases improved the classification performance compared with the baseline methods. The AUC's of each method were 0.90 (±0.04), 0.86 (±0.04), 0.81 (±0.05), and 0.74 (±0.05) respectively. The difference between the visual phrase method and the highest-performing baseline method (concatenated vector) was significant (p<0.05).

5 Conclusion and Discussion

In this work, we proposed a visual phrase learning scheme to learn a visual composite of anatomical parts from medical images for medical computer-aided detection.

In theory, components of a visual phrase are not necessarily "meaningful" anatomical parts. In the proposed method, useful visual words were identified by the learning process with the guidance of training labels. Any visual words which were useful for the classification were highly weighted and selected. For example, visual words on the colon folds and colon wall near the folds were highly weighted and therefore anatomically "meaningful". Experimental results on two CTC applications showed improved performance with the proposed method.

Our proposed method has several advantages. First, we do not need manual identification of the visual phrase. In Sadeghi and Farhadi's work on recognition using visual phrase [6], the appearance models for each category were learned using deformable part models which required manually labeled bounding boxes for training patches. Second, our method has more flexibility. We allow the learned visual parts to be distributed across the whole image, not limited to local patches.

Acknowledgements. This work was supported by the Intramural Research Programs of the NIH Clinical Center and by a Cooperative Research and Development Agreement with iCAD. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health.

References

1. Doi, K.: Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics 31, 198–211 (2007)
2. Wang, S.J., Summers, R.M.: Machine learning and radiology. Medical Image Analysis 16, 933–951 (2012)
3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision (2004)
4. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
6. Sadeghi, M.A., Farhadi, A.: Recognition Using Visual Phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2011)
7. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5, 27–72 (2004)
8. Cortes, C., Mohri, M., Rostamizadeh, A.: Learning sequence kernels. In: IEEE Workshop on Machine Learning for Signal Processing, pp. 2–8 (2008)
9. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proceedings of the Eighth IEEE International Conference on Computer Vision, ICCV 2001, vol. 521, pp. 525–531 (2001)