A BAG-OF-SHAPES DESCRIPTOR FOR MEDICAL IMAGING

Shuo Li, Andrea Lum, Gary Brahm, Ilanit Ben Nachum, Manas Sharma, Olga Shmuilovich, James Warrington

The digital imaging group of London
Dept. of Medical Imaging, Western University
London, ON, Canada

ABSTRACT

This paper proposes a new descriptor, Bag-of-shapes (BoS), to represent the global shape information in an image. The BoS descriptor is constructed by measuring the association likelihoods between an input image and a set of representative shapes learned off-line. Unlike existing global shape descriptors, BoS is not dependent on a segmentation of the input image, which makes it more practical for a heterogeneous environment like medical imaging. Other than image evidence alone, BoS also incorporates a high-level knowledge of expert shape preference, which makes it an enriched image representation. Furthermore, BoS has a naturally extensible form, i.e., multiple complementary BoS descriptors can be constructed for a single image, and then an optimal combination using a multi-kernel learning framework substantially strengthens the BoS descriptiveness. For validation, BoS is applied to a challenging task, cardiac cavity area estimation for both the left ventricle (LV) and the right ventricle (RV), which remains opening to the conventional segmentation techniques. BoS successfully tackles this task and produces highly consistent results with human expert on 76 clinical subjects. A comparison study also demonstrates that BoS outperforms six most popular existing peer descriptors.

Index Terms— Shape, Bag-of-Shapes, Medical Image, Cardiac

1. INTRODUCTION

Nowadays, more and more automatic analysis has been conducted by image based techniques. Within the process, a critical part is to represent an image effectively with a descriptor. Numerous descriptors have been proposed and can be categorized into two groups. (1) The first group reflects an image's global shape information, such as Moments, Curvature, Signature Histogram, Spectral Features, Shape Signature, Shape Context etc. [1] These descriptors are usually computed from the object boundary/region which is normally unavailable in the input image. Since an accurate object segmentation is extremely challenging in general, these shape descriptors are still immature and impractical for heterogeneous medical imaging applications. Readers interested in the existing shape representation and description techniques can refer to [1]. (2) Different from global shape descriptors, the second group of descriptors do not segment the input image and as expected they do not represent global shape information explicitly. Some of them capture the histogram based color/intensity or edge/gradient statistics in an image such as Edge Histogram Descriptor [2], HOG [3], WI-SIFT [4], and WI-SURF [4]. Some others are based on texture, such as GLCM [5] relying on the co-occurrence of gray values in different distances and orientations, BRIEF-GIST [6] relying on the construction of image models, and GIST [7] relying on signal processing results. Furthermore, there are descriptors based on the local appearance at salient points in an image such as BoW [8]. Most of the descriptors in the second group share an idea of representing an image with the collection of its grid or local patches' appearances, and this manner obviously does not represent global shape information.

In this paper, we propose a descriptor 'Bag-of-shapes' to represent an image's global shape information while without segmenting the input image. The idea is inspired by the BoW descriptor which represents the local shape information using a bag of visual words. Similarly, BoS takes advantage of a bag of shape words (a set of representative shapes learned off-line) to represent global shape information. Incorporated with a high-level shape prior, BoS is able to tackle the challenges encountered by appearance based descriptors, such as object intensity profile variation, gray level inhomogeneity, and surrounding structure disturbance. BoS is also naturally extensible in the sense that complementary BoS descriptors can be constructed for a single image and then optimally combined by a multi-kernel learning framework to further strengthen the descriptiveness. The BoS descriptor is validated by a challenging task, cardiac cavity area estimation for both LV and RV, which is critical for the diagnosis/treatment of cardiovascular diseases (the leading case of death [9]). The dramatic appearance variation and disturbing structures make this problem remain opening for LV and completely unsolved for RV to the conventional segmentation techniques [9]. With global shape information, BoS successfully tackles this problem and achieves a high conformity with human expert. A comparison study also demonstrates the superiority of BoS to six most popular existing peer descriptors.

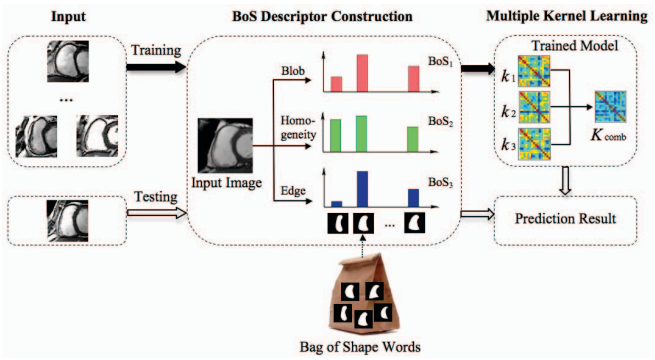


Fig. 1. Illustration of the BoS descriptor construction and its combination with multi-kernel learning framework.

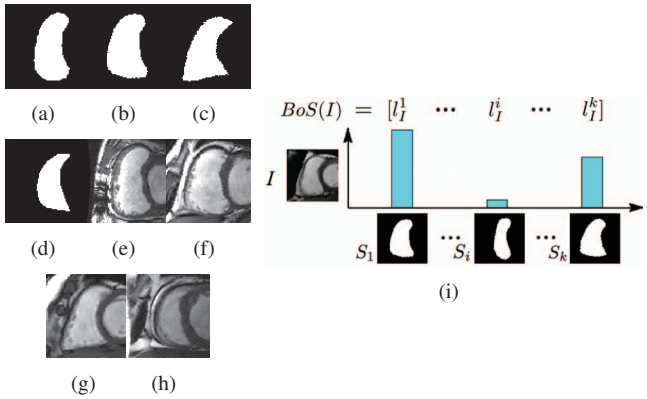


Fig. 2. (a-d) Shape words examples; (e-h) Corresponding original images where the shape words in (a-d) are segmented from by expert; (i) BoS descriptor illustration.

2. BAG-OF-SHAPES DESCRIPTOR

Similar to BoW which is formed by occurrences of a bag of visual words, a BoS descriptor, as shown in Fig. 1, is formed by the input image's association likelihoods to a bag of shape words measured by a selected feature. Multiple BoS descriptors are constructed by measuring the association likelihoods based on complementary features such as the examples (blob, homogeneity and edge) shown in the figure. The BoS descriptors are then fit into a multi-kernel learning framework seamlessly to perform prediction tasks. Note that all the examples used in this paper are based on RV images without loss of generality.

2.1. Shape words

The bag of shape words is a set of representative segmentation images constructed systematically off line:

– (1) A set of segmentation images, such as the experts' manual contouring results, are obtained from an auxiliary data set of the same application;

– (2) A k-medoids clustering [10] is conducted on the set of segmentation images;

– (3) The k representative cluster medoids are chosen and used as the bag of shape words.

Fig. 2(a-d) shows a few examples of shape words, obtained from experts' contouring results of Fig. 2(e-h). Incorporating contouring results to facilitate automatic analysis is common, such as (1) shape prior segmentation [11], (2) active shape and appearance models [12], (3) atlas-based segmentation techniques [13].

2.2. Descriptor construction

Given a bag of k shape words, a BoS descriptor of an image I can be defined as a k-dimensional vector (as shown in Fig. 2(i))

$$BoS(I) = [l_I^1, ..., l_I^i, ..., l_I^k], \quad (1)$$

where l_I^i is the normalized association likelihood between I and the i_{th} shape word S^i,

$$l_I^i = l(I, S^i) / \textstyle\sum_{j=1:k} l(I, S^j). \quad (2)$$

The association likelihood between an image I and a certain shape word S, $l(I, S)$, is measured by a feature response extracted from I in a way specifically determined by S as explained in the following. Various features can be used to measure this likelihood, and three examples, blob, homogeneity, and edge are discussed due to their efficient computation and popular existence in medical imaging.

2.2.1. *Blob feature based association likelihood*

The association likelihood $l(I, S)$ measured by blob feature can be computed as follows,

$$l_b(I, S) = (1 + \langle I, f^b(S) \rangle)/2 = (1 + \textstyle\sum_q I_q f_q^b(S))/2, \quad (3)$$

where $f^b(S)$ is an adaptive mask constructed specifically based on the shape word S. As shown in Fig. 3(a), this mask is composed of the inside (white) region, the surrounding band (black) region, and the remaining pixels. The pixel weights corresponding to the three regions sum to 1, -1, and 0 respectively, so that the dot product $\langle I, f^b(S) \rangle$ captures the image contrast in I between the regions corresponding to the inside and the surrounding band of S. The higher this contrast value is, the higher likelihood the object in I has a shape close to S, i.e., the bigger association between I and S.

2.2.2. *Homogeneity feature based association likelihood*

The association likelihood between I and S measured by homogeneity feature is computed as follows,

(a) (b) (c)

Fig. 3. The example adaptive masks used in the three association likelihood functions (taking the RV case as example). (a) Blob mask captures the image contrast by computing its dot product with the input image; (b) homogeneity mask specifies the cavity region within which the homogeneity is computed; (c) edge mask captures the edge feature by computing its dot product with an edge image.

$$l_h(I,S) = 1 - \sum_{q\in\Omega_c(S)} \left(I_q - \mu(I,S)\right)^2 / |\Omega_c(S)|, \quad (4)$$

$$\mu(I,S) = \sum_{q\in\Omega_c(S)} I_q / |\Omega_c(S)|. \quad (5)$$

$\mu(I,S)$ is the mean intensity of I within the region of S, as shown by the example mask in Fig. 3(b). $|\Omega_c(S)|$ is the area of shape region $\Omega_c(S)$. The more homogeneous I is within the region S, the higher value $l_h(I,S)$ is, and the bigger association between I and S is.

2.2.3. Edge feature based association likelihood

The association likelihood between I and S measured by edge feature is computed as follows,

$$l_e(I,S) = \langle I^e, f^e(S)\rangle = \sum_q I^e_q \cdot f^e_q(S), \quad (6)$$

where $f_e(S)$ is the adaptive edge mask (as the example in Fig. 3(c)) constructed based on S. The mask is composed of the boundary of S with pixel weights sum to 1 and the rest part with pixel weights 0. The edge images is computed by $I^e = \sqrt{(\partial_r I)^2 + (\partial_c I)^2}$ where $\partial_r I$ and $\partial_c I$ are the first derivatives in the row and column directions respectively.

3. MULTIPLE KERNEL LEARNING OF BOS DESCRIPTORS

Multiple BoS descriptors can be constructed for an image based on complementary features, and a combination of them is expected to provide strengthened descriptiveness. Hand-tuning this combination is difficult and ineffective, and therefore the advantage of multi-kernel learning is leveraged in this paper to learn an optimal descriptor combination from training data. The learning method is adopted from [14] due to its ability to learn general kernel combinations for far richer representations than linear format.

Following the format of the product of RBF (radial basis function) kernels, we define the combined kernel function of two images (in vector form) as follows,

$$K_{\mathbf{d}}(I,I') = \Pi_m \exp(-\mathbf{d}_m \cdot ||BoS_m(I) - BoS_m(I')||^2), \quad (7)$$

where each RBF kernel is specifically constructed based on one BoS descriptor and $\mathbf{d}$ is the parameter vector weighting these RBF kernels. The objective of a regression or classification task using Support Vector Machine (SVM) is to learn a function of the form

$$f(I) = \mathbf{w}^t \phi_{\mathbf{d}}(I) + b \quad (8)$$

to map an image I to a predict value. The kernel $K_{\mathbf{d}}(I,I') = \phi^t_{\mathbf{d}}(I)\phi_{\mathbf{d}}(I')$ represents the dot product in feature space ϕ parameterized by $\mathbf{d}$. Learning the function $f(I)$ is fulfilled by solving the following optimization problem,

$$\begin{array}{ll} \text{Min}_{\mathbf{w},b,\mathbf{d}} & \frac{1}{2}\mathbf{w}^t\mathbf{w} + \sum_i l(y_i, f(I_i)) + r(\mathbf{d}) \\ \text{subject to} & \mathbf{d} \geq 0. \end{array} \quad (9)$$

where r is a differentiable regularizer function of $\mathbf{d}$ and l is a loss function. $\{(I_i, y_i)\}$ is the training set associating each input I_i with a predict value y_i. The above optimization problem is solved by a standard nested two step optimization procedure formally represented as

$$\begin{array}{ll} \text{Min}_{\mathbf{d}} & T(\mathbf{d}) \quad \text{subject to } \mathbf{d} \geq 0 \\ \text{where} & T(\mathbf{d}) = Min_{\mathbf{w},b}\frac{1}{2}\mathbf{w}^t\mathbf{w} + \sum_i l(y_i, f(I_i)) + r(\mathbf{d}). \end{array} \quad (10)$$

More detailed solution to Eq. (10) can be referred to [14].

The seamless combination of BoS and multi-kernel learning boosts each other's strength and enables us to tackle the challenging application employed in this paper, cardiac cavity area estimation, which remains opening to the conventional segmentation techniques.

4. EXPERIMENTAL EVALUATION

In this experiment, we will demonstrate the feasibility of the BoS descriptor to the cardiac cavity area estimation problem as well as its superiority to six most popular existing peer descriptors, HOG, BRIEF-GIST, GIST, WI-SIFT, WI-SURF and BoW. Two datasets were used in the experiment,

- (1) The LV dataset contains 3000 2-D short-axis cine MR images (80 pixel by 80 pixel) of 50 clinical subjects from the public dataset, LV-challenge 2009 [15]. Additional 2400 contoured LV from the LV-challenge dataset were used to construct the bag of shape words;
- (2) The RV dataset contains 1560 two-dimensional (2-D) short-axis cine MR images (80 pixel by 80 pixel) collected from 26 clinical subjects at the authors' associated hospital. Additional 1800 contoured RV from similar data were used to construct the bag of shape words.

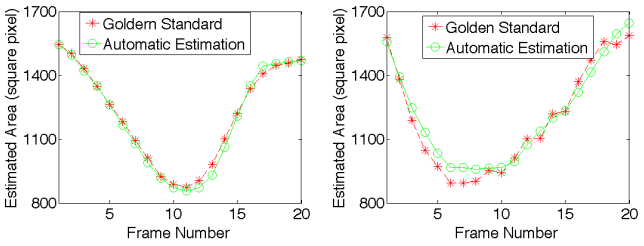
The test results on both the LV and RV datasets were obtained based on a *leave-one-out* strategy. Three metrics were employed for evaluation purpose: absolute error ($Error_{abs}$), relative error ($Error_{rel}$), and correlation coefficient ($corr(Area_A, Area_M)$) between the automatically

Table 1. Comparison between the proposed BoS descriptor and six most popular existing descriptors: HOG, BRIEF-GIST, GIST, WI-SIFT, WI-SURF, and BoW on the LV dataset.

Method	*mean*($Error_{abs}$)	*mean*($Error_{rel}$)	*Corr* ($Area_A$,$Area_M$)
BoS	**85.1±69.1**	**6.96%**	**0.9392**
HOG	110.1±84.5	9.27%	0.9006
BRIEF-GIST	148.5±119.0	12.11%	0.8058
GIST	117.8±93.4	9.86%	0.8813
WI-SIFT	161.3±121.6	13.32%	0.7836
WI-SURF	156.8±131.2	13.09%	0.7747
BoW	204.8±157.4	17.18%	0.5955

Table 2. Comparison between the proposed BoS descriptor and six most popular existing descriptors: HOG, BRIEF-GIST, GIST, WI-SIFT, WI-SURF, and BoW on the RV dataset.

Method	*mean*($Error_{abs}$)	*mean*($Error_{rel}$)	*Corr* ($Area_A$,$Area_M$)
BoS	**121.6±101.8**	**11.32%**	**0.9141**
HOG	176.0±145.9	16.88%	0.8049
BRIEF-GIST	227.2±166.4	21.35%	0.6838
GIST	265.0±190.1	25.41%	0.6704
WI-SIFT	215.0±174.7	19.95%	0.7353
WI-SURF	192.0±140.8	18.01%	0.7909
BoW	265.6±215.0	27.12%	0.4825



(a) LV (b) RV

Fig. 4. Illustration of the high conformity between the automatically estimated areas and the golden standard of an example LV subject (a) and an example RV subject (b) over a full cardiac cycle.

estimated cavity areas ($Area_A$) and the manually obtained golden standard areas ($Area_M$).

The BoS descriptor produced highly consistent results with human expert on both LV and RV as shown in Fig. 4 by the two example subjects over a full cardiac cycle. The high conformity is also emphasized in Table 1 and 2 by a low absolute error 85.1/121.6 square pixels, a low relative error 6.96%/11.32%, and a high correlation coefficient 0.9392/0.9107 on LV/RV. The further comparison between the BoS descriptor and the six competitor descriptors on both LV and RV are also reported in Table 1 and 2. For both cases, the BoS descriptor produced much better results than the six competitors, which clearly demonstrated the benefit of global shape information from the shape words. Specifically in the RV case, this benefit is more obviously illustrated by a remarkable correlation coefficient gap of more than 0.1 between the BoS result (0.9141) and the best competitor result (0.8049). This is because the RV has a more complicated appearance and surrounding background than LV, e.g., low contrast between the blood pool and the myocardium, disturbance from surrounding tissues such as fat, and the existence of trabeculations inside the RV chamber. As a result, the existing descriptors based on appearance are substantially affected and cannot produce results as accurate as BoS.

5. CONCLUSION

This paper proposed the BoS descriptor which is able to represent an image's global shape information without segmenting the input image. The BoS descriptor was tested in an application with significant clinical importance, cardiac cavity area estimation for both LV and RV. The experimental results conducted on two data sets of 50 and 26 clinical subjects showed that the proposed descriptor produced highly consistent results with human expert and also outperformed six most popular existing peer descriptors. Indeed, the idea of estimating the object area by BoS along with multi-kernel learning is rather flexible and can be potentially applied to other modalities and/or problems where the object area information needs to be estimated.

6. REFERENCES

[1] Zhang, D., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition **37**(1) (January 2004) 1–19

[2] Annesley, J., Orwell, J., Renno, J.P.: Evaluation of MPEG7 Color Descriptors for Visual Surveillance Retrieval. 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005) 105–112

[3] Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) **1** (2005) 886–893

[4] Badino, H., Huber, D., Kanade, T.: Real-time Topometric Localization. 2012 IEEE International Conference on Robotics and Automation (May 2012) 1635–1642

[5] Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics **3**(6) (November 1973) 610–621

[6] Sunderhauf, N., Protzel, P.: BRIEF-Gist - Closing the Loop by Simple Means. 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (September 2011) 1234–1241

[7] Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision **42**(3) (2001) 145–175

[8] Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. Proceedings Ninth IEEE International Conference on Computer Vision **2** (2003) 1470–1477

[9] Petitjean, C., Dacher, J.N.: A Review of Segmentation Methods in Short Axis Cardiac MR Images. Medical Image Analysis **15**(2) (April 2011) 169–184

[10] Kaufman, L., Rousseeuw, P.: Clustering by Means of Medoids. In Dodge, Y., ed.: Statistical data analysis based on the L-1 norm. Elsevier, Amsterdam (1987) 405–416

[11] Cremers, D., Rousson, M., Deriche, R.: A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape. International Journal of Computer Vision **72**(2) (August 2006) 195–215

[12] Heimann, T., Meinzer, H.P.: Statistical Shape Models for 3D Medical Image Segmentation: A Review. Medical image analysis **13**(4) (August 2009) 543–63

[13] Rohlfing, T., Brandt, R., Menzel, R.: Quo Vadis, Atlas-Based Segmentation? In: Handbook of Biomedical Image Analysis. (2005)

[14] Varma, M., Babu, B.R.: More Generality in Efficient Multiple Kernel Learning. Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09 (2009) 1–8

[15] Sunnybrook Health Sciences Centre: Cardiac MR Left Ventricle Segmentation Challenge. http://smial.sri.utoronto.ca/LV_Challenge/Home.html (2009)