Medical Image Annotation and Retrieval Using Visual Features

Jing Liu[1]*, Yang Hu[2]*, Mingjing Li[3], and Wei-ying Ma[3]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
jliu@nlpr.ia.ac.cn
[2]University of Science and Technology of China, Hefei 230027, China
yanghu@ustc.edu
[3]Microsoft Research Asia, No 49, Zhichun Road, Beijing 100080, China
{mjli, wyma}@microsoft.com

Abstract

In this article, we present the algorithms and results of our participation in the medical image annotation and retrieval tasks of ImageCLEFmed 2006. We exploit both global features and local features to describe medical images in the annotation task. We examine different kinds global features and extract the most descriptive ones, which effectively capture the intensity, texture and shape characters of the image content, to represent the radiographs. We also evaluate the descriptive power of local features, i.e. local image patches, for medical images. A newly developed spatial pyramid matching algorithm is applied to measure the similarity between images represented by sets of local features. Both descriptors use multi-class SVM to classify the images. The error rate is 17.6% for global description and 18.2% for the local one, which rank sixth and ninth respectively among all the submissions. For the medical image retrieval task, we only use visual features to describe the images. No textual information is considered. Different features are used to describe gray images and color images. Our submission achieves a mean average precision (MAP) of 0.0681, which ranks second in the 11 runs that also only use visual features.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image annotation, Image retrieval, Support vector machine, Similarity measure

[*] This work was performed when the first and the second authors were visiting students at Microsoft Research Asia.

1 Introduction

Due to the rapid development of biomedical informatics, medical images have become an indispensable investigation tool for medical diagnosis and therapy. A single average size radiology department may produce tens of tera-bytes of data annually. The ever-increasing amount of digitally produced images require efficient methods to archive and access this data. Therefore, the application of general image classification and retrieval techniques in this specialized domain has obtained increasing research interest recently.

ImageCLEF, which conducts evaluation of cross-language image retrieval has come up with a medical image retrieval task since 2004. And an automatic medical image annotation task was added in 2005. It provides a benchmark to evaluate the performance of different algorithms on the same tasks using the same dataset. The tasks in 2006 are similar to those in the last year. The dataset and the task description are almost the same. However, the topics are more challenging than last year's. More categories are defined for the annotation task and more semantic queries are issued for the retrieval task.

In this paper, we describe our participation in the automatic medical image annotation and medical image retrieval tasks of ImageCLEF 2006. We submitted two runs for the annotation task, which exploited the effectiveness of two different kinds of features to describe and classify medical images. The first run examined different kinds of global features and extracted the most descriptive ones to represent the radiographs. It achieved an error rate of 17.6%, which ranked sixth among all the submissions. In the second run, we applied a newly developed spatial pyramid matching scheme to this task, which effectively measured the similarity between images represented by sets of local features. It achieved an error rate of 18.2%, and ranked ninth in the submissions. We submitted one run for the medical image retrieval task. We evaluated the effectiveness of visual features for medical image retrieval. Our submission yielded a mean average precision (MAP) of 0.0681, which ranked second in the 11 runs that also only used visual features.

The rest of the paper is organized as follows. We describe the details of our runs for the automatic annotation task in Section 2. The medical image retrieval task is presented in Section 3. Experimental results are discussed in Section 4. Finally, we conclude this paper in Section 5.

2 Automatic Medical Image Annotation

The automatic image annotation task is to classify images into a set of predefined categories. It provides a dataset consisting of 10,000 fully classified radiographs for participants to train a classification system. These images are classified into 116 categories this year according to image modality, body orientation, body region and the biological system examined. 1000 additional radiographs whose classification labels are unavailable to participants are used to evaluate the performance of various algorithms.

We developed two different schemes for this task. In the first algorithm, traditional global features, such as intensity, texture and shape descriptors were used to describe medical images. In the second one, we exploited using local features to represent the images. And a spatial pyramid matching scheme was then applied to measure the similarity between two images. Both methods used SVM to classify the images into different categories.

2.1 Global Features for Medical Image Classification

When designing image features, we should consider two issues. First, the features should be representative for the images. Second, the complexity of calculating the features should be relatively low. Medical images have their particular characteristics in appearance. For example, radiographs are usually grayscale images and the spatial layouts of the anatomical structures in the radiographs of the same category are quite similar. The texture, shape and local features are valuable and discriminative for describing medical images.

According to these observations, we select several different visual features to represent the radiographs. We extract gray-block feature and block wavelet feature from the original images. Shape-related features are exacted from the corresponding binary images. Then, they are combined into a 382-dimensional feature vector. The detail descriptions of the features are as follows:

Gray-block feature The original images are uniformly divided into $8 \times 8 = 64$ blocks. The average gray value in each block is calculated and a 64-dimensional gray-block feature is obtained. The ℓ_2-norm of the feature vector is set to 1. The normalization could reduce the influence of illumination variance across different images to some extent. According to the experiments, this is the most effective feature although it is straight forward and very simple.

Block-wavelet feature The wavelet coefficients could characterize the texture of the images at different scales. We divide the images into $4 \times 4 = 16$ blocks and extract multi-scale wavelet features in each block. We implement 3-level wavelet transforms on the image blocks using Daubechies filter (db8). Then, the mean and the variance of the wavelet coefficients in the HL, LH and HH sub-bands are computed. Therefore, we get a $288(6 \times 3 \times 4 \times 4)$-dimensional feature vector.

Features for the binary image We first convert the images into binary images. Otsu's method [10] is used here to calculate the threshold. The area and the center point of the object region in the binary image are calculated. Moreover, we apply morphological operations on the binary image and extract the contour and the edges of the image. The length of the contour and the ratio of the total length of the edges and that of the contour are calculated and are taken as the shape feature. Then we get a 5-dimensional feature for the binary image. Although the dimension of this feature is small, it is highly discriminative among different categories. In order to increase the effect of this feature, we duplicate it 6 times and convert it into a 30-dimensional feature vector.

Choosing suitable parameters for above features is very difficult in theory. Therefore, we tune the parameters through experiments. The parameters, such as the size of the image block and the dimension of the features for the binary image, are determined through cross-validation on the training set. The same parameter settings are used in both of the annotation task and the retrieval task.

The classifier is trained using SVM, which is a classic machine learning technique that has strong theoretical foundation and excellent empirical successes. The basic idea of SVM is to map the data into a high dimensional space and then find a separating hyperplane with the maximal margin. In the experiment, we use the multi-class SVM implemented by the LIBSVM tool[9]. The radial basis function (RBF) is chosen as the kernel function and the optimal parameters are determined through 5-fold cross-validation.

2.2 Spatial Pyramid Matching for Medical Image Classification

Recently, a class of local descriptor based methods, which represent an image with an collection of local photometric descriptors, have demonstrated impressive level of performance for object recognition and classification. And this kind of algorithms have also been explored for medical image classification, considering that most information in medical images is local [1]. Unlike global features, local features are always unordered. Different images are represented by different number of local descriptors and the correspondence between the features across different images is unknown. Therefore, it is challenging to apply this kind of representation to discriminative learning, which usually operates on fixed-length vector inputs. Many recent works have devoted to leverage the power of both local descriptor and discriminative learning. In [2], Grauman and Darrell proposed to map sets of features to multi-resolution histogram and then compare the histograms with a weighted histogram intersection measure. The pyramid matching scheme resulted in a kernel which was proved to satisfy Mercer's condition. And SVM was then trained to recognize the objects. Inspired by the idea of [2], Lazebnik et al.[3] presented a spatial pyramid matching method for recognizing natural scene categories. Instead of exploiting the structure of feature space, it constructed pyramid in image space by partitioning the image into increasingly fine sub-regions. The histograms of local features were computed on each sub-region and the same weighted histogram intersection was applied to measure the similarity between feature sets.The geometric information of local features is extremely valuable for medical images, since the objects are always centered in the images and the spatial layouts of the anatomical structures in the radiographs belonging to the same category are quite similar. Therefore, we can expect promising results using this spatial matching scheme. We apply spatial pyramid matching for medical image classification and examine its performance on this new task.

Although SIFT descriptor [4] has been proven to work well for common object and nature scene recognition [2][3], its power to describe radiographs is somewhat limited. Since the scale and rotation variations in radiographs of the same category are small, the SIFT descriptor can not show its advantage of being scale and rotation invariant for describing radiographs. In previous works, local image patches have shown pleasant performance for medical image retrieval and classification [5][6][7]. Therefore, we utilize local image patches as the local features in our experiments. Before feature extraction, we resize the images so that the long sides are 200 pixels and their aspect ratios are maintained. The positions of the local patches are determined in two ways. Local patches are first extracted from interest points detected by DoG region detector [4], which are located at local scale-space maxima of the Difference-of-Gaussian. We also extract local patches from an uniform grid spacing at 10 × 10 pixels. This dense regular description is necessary to capture uniform regions that are prevalent in radiographs. We use 11 × 11 pixel patches in our experiments. And about 400 patches are extracted from each image. After feature extraction, we applied a high speed clustering algorithm Growing Cell Structures (GCS) neural network [8], which is able to detect high dimensional patterns with any probability distribution, to quantize all feature vectors into M discrete types ($M = 600$ in the experiment). Then each feature vector is represented by the ID of the cluster it belongs to and its spatial coordinate.

In order to measure the similarity between two images represented by orderless collections of local patches, we first partition the scaled images into increasingly fine sub-regions. Then we compute the histograms of cluster frequencies inside each sub-region by counting the number of patches that belong to each cluster (Fig. 1). The histograms from two images are compared using a weighted histogram intersection measure. Let X and Y be two sets of feature vectors representing two images. Their histograms in the ith sub-region at level l are denoted by H_X^{li} and H_Y^{li} with $H_X^{li}(j)$ and $H_Y^{li}(j)$ indicating the number of feature vectors from X and Y that fall into the jth bin of the histograms. The histogram intersection function is given by

$$\mathcal{I}(H_X^{li}, H_Y^{li}) = \sum_{j=1}^{M} \min(H_X^{li}(j), H_Y^{li}(j)) \ , \tag{1}$$

which measures the "overlap" between two histograms' bins. It implicitly finds the correspondences between feature vectors falling into that sub-region. The similarity between X and Y is defined as the weighted sum of the number of matches found in each sub-region:

$$\mathcal{K}(X, Y) = \sum_{l=1}^{L} w_l \sum_{i=1}^{4^{(l-1)}} \sum_{j=1}^{M} \min(H_X^{li}(j), H_Y^{li}(j)) \ , \tag{2}$$

where L refers to the max level. As shown in Fig. 1, the weight w_l is inversely proportional to region size: the smaller the region the larger the weight, i.e. matches made within smaller regions are weighted more than those made in larger regions.

Actually, $\mathcal{K}$ can be implemented as a single histogram intersection of "long" vectors which are formed by concatenating the appropriately weighted histograms in all sub-regions. For L levels and M clusters, although the index of the single histogram may be as high as $M \sum_{l=1}^{L} 4^{l-1}$, the histogram of each image is actually very sparse. The number of non-zero bins is at most mL.

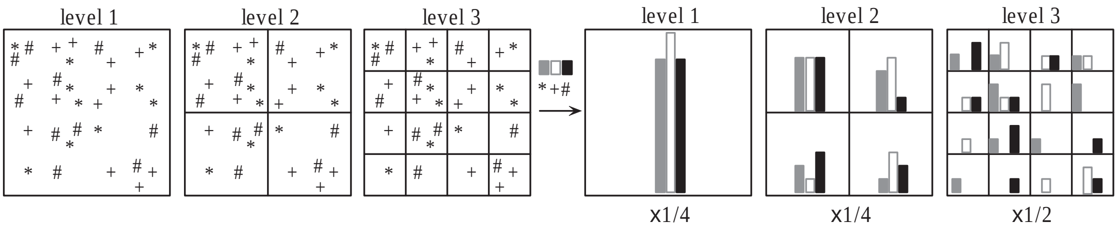

Figure 1: Toy example of constructing a three-level spatial pyramid. The image has three types of features, indicated by asterisks, crosses and pounds. At the left side, the image is subdivided at three different levels of resolution. At the right, the number of features that fall in each sub-region is counted. The spatial histograms are weighted during matching [3].

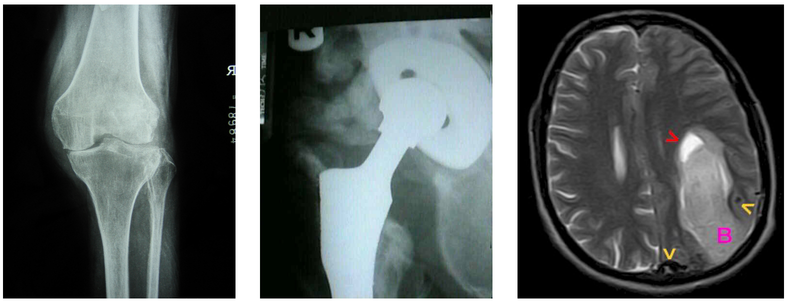

Figure 2: Example query images which are regarded as gray images.

Another implementation issue is normalization. In order not to favor large feature sets, which would always yield high similarity due to the intersection operation, we should normalize the histograms by the total weight of all features in the images before conducting matching.

$\mathcal{K}$ has been proved to satisfy the Mercer's condition, i.e. it is positive semi-definite [2][3]. Therefore, kernel-based discriminative methods can be applied. In the experiment, multi-class classification is done with a "one-against-one" SVM classifier [9] using the spatial pyramid matching kernel.

3 Medical Image Retrieval

The dataset for the medical image retrieval task consists of images from the Casimage, MIR, PEIR and PathoPIC datasets. There are totally 50,026 images with different modalities, such as photographs, radiographs, ultrasonic images, and scans of illustrations used for teaching etc. Query topics are formulated with example images and a short textual description, which denotes the exact information need such as the illness, the body region or the modalities shown in the images. Therefore, this task is much more challenging than the annotaion task. We only exploit the effectiveness of visual features for this task. No textual information is utilized in our experiment.

As general image retrieval systems, the whole retrieval procedure contains three steps: image preprocessing, feature extraction and relevance ranking based on similarity measure. For image preprocessing, we first resize the images so that the long sides are 512 pixels and their aspect ratios are maintained. As the characters of gray images and color images are quite different, we examine whether an image is gray or color before extracting features from it. Note that the images in Fig. 2 are regarded as gray images because the color information in them are very limited and also useless for retrieval. Feature extraction is carried out according to the type of the image, i.e. the features for gray image and color image are different:

Features for gray images The global features used to describe radiographs in the annotation

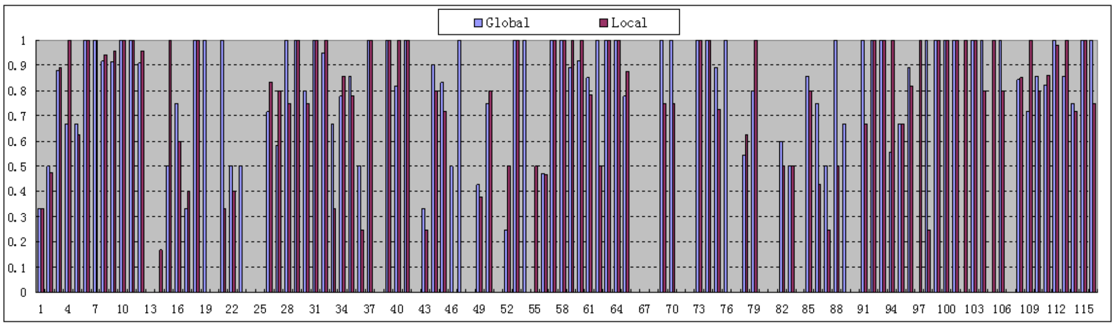

Figure 3: Classification precisions for each category on the test set.

task are used here to describe the gray images.

Features for color images We use band-correlogram, color histogram and block-wavelet features to describe the color images:

- *Band-correlogram* We first quantize the RGB values into 64 bins. Then the general autocorrelogram features are extracted within four square neighborhoods, whose radius are 1,3,5,7 pixels respectively. The final features used are the average of the corresponding elements in the four square neighborhoods. It is a 64-dimensional feature vector.
- *Color histogram* We quantize the RGB values into 36 bins, and calculate the 36-dimensional color histogram as described in [11].
- *Block-wavelet* We first convert the color images into gray images using:

$$L = 0.299 \times R + 0.587 \times G + 0.114 \times B \ . \tag{3}$$

Then the block-wavelet feature are calculated as introduced in Sect.2.1.

The last step is ranking the images in the dataset according to their relevance to the query images. As each topic contains multiple query images, the distance between a dataset image Z and a set of query images belonging to the same topic is defined as the minimun distance between Z and each query image:

$$d(Z, Q) = \min_i d(Z, Q_i) \ . \tag{4}$$

The top 1000 images are returned for evaluation.

4 Experimental Results

4.1 Results of Automatic Medical Image Annotation

For the annotation task ,we submitted two runs named “msra_wsm_gray” and “msra_wsm_patch” for global feature and local feature methods respectively. The submission using global features achieved an error rate of 17.6%, which ranked sixth among all the submissions. And the error rate of the run using local features is 18.2%, which ranked ninth.

Fig. 3 illustrates the classification precisions of each category on the test dataset. The results for the run using global features are denoted by blue bars, and the local feature based method is denoted by red bars. In Fig. 4 we calculate the average precisions across different categories, for which the numbers of training images are larger than a specified number given by the X axis. Through analyzing these experimental results, we could get some valuable information. Firstly, all the categories with zero precisions are corresponded to the categories whose training images are less than 20. Secondly, when the number of training images is larger than 20, our

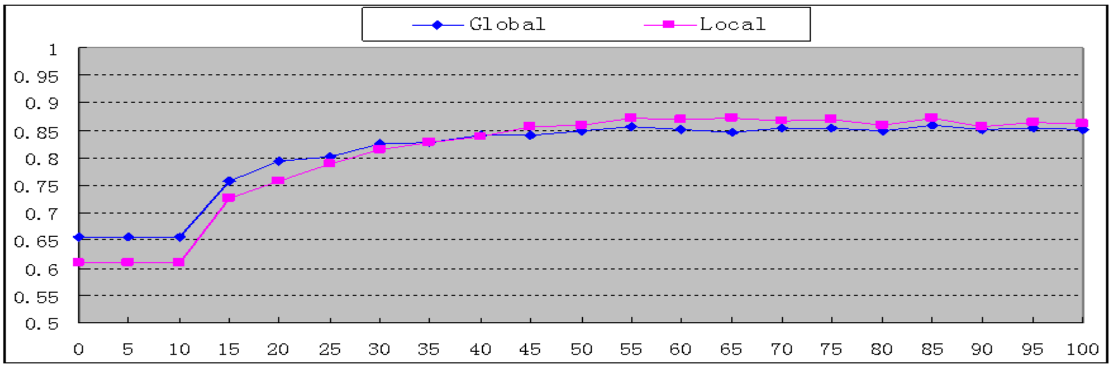

Figure 4: Average precisions across categories, for which the numbers of training images are larger than the number specified by the X axis.

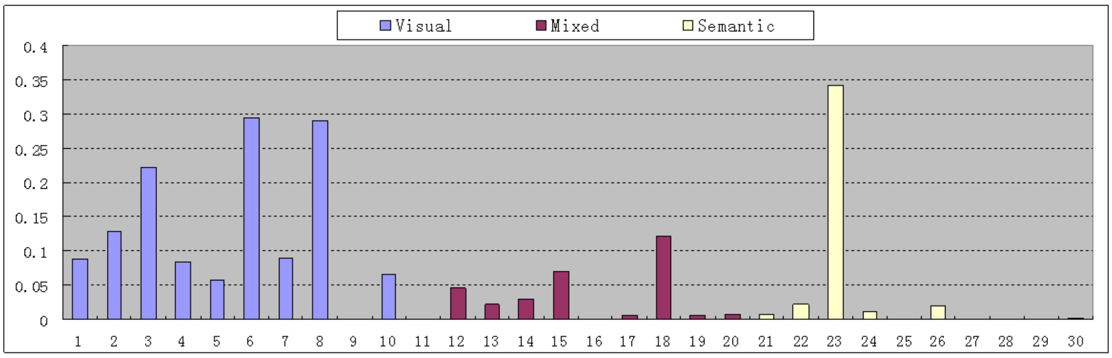

Figure 5: Mean average precision per query topic.

methods could have more stable performance on average precision. Thirdly, our two methods achieved comparable performances. As they are complementary for describing images, we could expect better performance if we combine these two descriptions together. However, we haven't implemented the combination so far. We will explore it in our future work.

4.2 Results of Medical Image Retrieval

In the medical image retrieval task, the parameters for gray images are the same with the annotation task. The parameters for color images are determined empirically. The details have been discussed in Section 3. We employ these features in our automatic "visual retrieval" system and submit only one run named "msra_wsm". We achieved a MAP of 0.0681, which ranks second among the 11 runs that also only use visual features. The MAP of the best run is 0.0753.

The MAP values for each query are shown in Fig. 5. We use different color bars to indicate the different performances on visual, mixed and semantic topics. The average MAP on these three kinds of topics are 0.1324, 0.0313 and 0.0406 respectively. It is obvious that the performance on visual topics is the best. The performance is relatively poor on other topics with more semantic considerations. The differences between the performances on different kinds of topics are reasonable considering the design of the topics. The MAP for the 23rd topic which is a semantic topic is strangely high. It is because the number of images that are similar with the query images of this topic is quite large.

5 Conclusion

In this paper, we present our work on the medical image annotation and retrieval tasks of ImageCLEFmed 2006. Due to the special characteristics of medical images, we explored using global and local features respectively to describe the radiographs in the annotation task. Then we use the multi-class SVM to classify the images. We achieved an error rate of 17.6% for the global feature based method and 18.2% for the local feature method. For the medical image retrieval task, we distinguished gray images from color images and used different kinds of visual features to describe them. Our submission ranked second among the 11 runs which also only used visual features.

This is our first participation in the tasks concerning medical images. We find this task quite interesting and very challenging. In our future work, we will investigate some more descriptive features and more suitable similarity measure for comparing images. We didn't utilize the textual information in our experiment. We will incorporate it into the retrieval framework in the future.

References

[1] Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., Wein, B.B.: Automatic Categorization of Medical Images for Content-based Retrieval and Data Mining. Computerized Medical Imaging and Graphics, volume 29, pages 143-155, , 2005.

[2] Grauman, K., Darrell, T.: The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. Proceedings of the IEEE International Conference on Computer Vision (ICCV 2005), Beijing, China, October 2005.

[3] Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, June 2006.

[4] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[5] Keysers, D., Gollan, C., Ney, H.: Classification of Medical Images using Non-linear Distortion Models. Bildverarbeitung für die Medizin 2004 (BVM 2004), Berlin, Germany, pages 366-370, March 2004.

[6] Deselaers, T., Keysers, D., Ney, H.: Discriminative Training for Object Recognition Using Image Patches. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, June 2005.

[7] Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Biomedical Image Classification with Random Subwindows and Decision Trees. Proceedings of ICCV workshop on Computer Vision for Biomedical Image Applications (CVIBA 2005), Beijing, China, October 2005.

[8] Fritzke, B.:Growing Cell Structures – A Self-Organizing Network in k Dimensions. Artificial Neural Networks II, pages 1051-1056, 1992.

[9] Chang, C.-C., Lin, C.-J.: LIBSVM : A Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm .

[10] Otsu, N.: A Threshold Selection Method from Gray-Level Histogram. IEEE Trans. System Man Cybernetics, SMC-9(1): 62-66, 1979.

[11] Swain, M. and Ballard, D.: Color Indexing. International Journal of Computer Vision, Vol. 7, No. 1, 1991.