Approximated Classification in Interactive Facial Image Retrieval

Zhirong Yang and Jorma Laaksonen

Laboratory of Computer and Information Science,
Helsinki University of Technology,
P.O. Box 5400, FI-02015 HUT, Espoo, Finland
{zhirong.yang, jorma.laaksonen}@hut.fi

Abstract. For databases of facial images, where each subject is depicted in only one or a few images, the query precision of interactive retrieval suffers from the problem of extremely small class sizes. A potential way to address this problem is to employ automatic even though imperfect classification on the images according to some high level concepts. In this paper we point out that significant improvement in terms of the occurrence of the first subject hit is feasible only when the classifiers are of sufficient accuracy. In this work Support Vector Machines (SVMs) are incorporated in order to obtain high accuracy for classifying the imbalanced data. We also propose an automatic method to choose the penalty factor of training error and the width parameter of the radial basis function used in training the SVM classifiers. More significant improvement in the speed of retrieval is feasible with small classes than with larger ones. The results of our experiments suggest that the first subject hit can be obtained two to five times faster for semantic classes such as "black persons" or "eyeglass-wearing persons".

1 Introduction

Most existing face recognition systems require the user to provide a starting image. This however is not practical in some situations, e.g., when searching a previously seen image via the user's recalling. To address this problem, some interactive facial image retrieval systems such as [9] have been proposed, which are mainly based on learning the relevance feedback from the user.

Unlike content-based image retrieval (CBIR) systems on general images, the query precision on facial images suffers from the problem of extremely small class sizes [9]. In a popular collection, FERET [7], most subjects possess only one or two frontal images. Making the *first subject hit* appear as early as possible is critical for the success of interactive facial image retrieval. If only images that depict the correct person are regarded as relevant, many zero pages (i.e. the images in these rounds are all non-relevant) will be displayed until the first relevant image emerges. This is because the negative responses from the user in early rounds provide only little semantic information and – as a result – the iteration progresses in a nearly random manner.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 770–779, 2005.

The above problem can be relieved by allowing the user to submit partial knowledge, e.g. gender or race, on the query target. With this kind of *restriction* or *filtering* there are far less image candidates than the entire collection and the first subject hit will undoubtedly appear much sooner. However, this requires labeling of the images according to the semantic criteria and manual work is not feasible for a large database. Thus approximating the semantic annotation by automatic classification is desired.

Classifiers constructed by machine learning are generally not perfect. If the correct target happened to be misclassified then it would never be displayed due to the filtering. In this paper we suppose the user would not give up an unsuccessful query after some number of endurable rounds – instead he or she would remove the restriction and continue the search by using the entire database. This assumption allows us to compute the mean position of the first subject hit and assess the advantage obtainable with approximated classification.

In this paper we point out that only classifiers with very high accuracy can be significantly beneficial to the retrieval. This basic assumption is verified by experiments in Section 2. Support Vector Machines are used for automatic classification. We review SVM's principles and discuss how to choose its parameters with radial basis function kernels in Section 3. Experiments are presented in Section 4, and finally are the conclusions and future work in Section 5.

2 Approximated Classification

Restricting the image candidates by some *true semantic classes* is a natural idea to improve query performance. However in CBIR the true semantic classes are usually not available and we have to approximate them by *restriction classes* which can be defined by some automatic classifiers. If a classifier constructed by machine learning has only a small misclassification error rate, the first relevant image can be shown earlier on the average. In this section we will present a set of preliminary experiments to sustain the idea.

2.1 First Subject Hit Advantage Performance Measure

The position of the first relevant hit is critical to the success of CBIR. For example, if there is no relevant image displayed in the first twenty rounds, the user would probably deem the system useless and give up the query. In contrast, if the first relevant hit appears within the user's tolerance, say the first five or ten rounds, the query will probably proceed and further relevant images found.

Suppose N and R are the number of all images and relevant images in the database, respectively. Denote by j the random variable for position of the first subject hit using random retrieval. It is not difficult to prove that the mean of j is $E\{j\} = (N - R)/(R + 1)$. Thus the improvement compared with the random retrieval can be quantified by the following *first subject hit advantage* (FSHA) measurement:

$$\mathrm{FSHA}(i; N, R) = \frac{E\{j\}}{i} = \frac{N - R}{i \cdot (R + 1)}, \tag{1}$$

where $i \in \{0, 1, \ldots, N - R\}$ is the position of the first subject hit using the improved retrieval. FSHA equals one when the retrieval is done in a random manner and increases when the retrieval is able to return the first relevant image earlier. For example, it equals two when the first subject hit occurs in the position whose index is half of the expected index in random retrieval.

2.2 Simulated Classification Study

In order to test the advantage attainable by queries with approximated classification, our experiments were carried out as follows. For a set of semantic classes $\{C_i\}$, we simulated approximated classification by random sampling with varying percentages of correct and incorrect decisions according to the criterion C_i. For each simulated class and classification accuracy we then ran the PicSOM CBIR system [5] to calculate the attained average FSHA. We looped over all subjects $\{S_t\}$ in the class C_i and at each loop, the retrieval goal was to search all images depicting the current subject S_t. 20 images were "displayed" per round and the first set of images was randomly selected from C_i. In the automated evaluation the sole criterion for relevance of an image was whether it depicted the current subject S_t or not. If no subject hit appeared within a predefined number of rounds, T, the target was deemed to have been misclassified. The test program then removed the restriction and resorted to using the entire database until the first subject hit occurred.

In the experiments we used the FERET database of readily segmented facial images collected under the FERET program [7]. 2409 frontal facial images (pose mark "fa" or "fb") of 867 subjects were stored in the database for the experiments. Table 1 shows the specification of four tested true semantic classes.

We used two different T values, 10 and 20, to study the query performance. The results are shown in Figure 1. The FSHAs using the entire database are shown as the dotted baseline at the bottom of the plots. Due to the extremely small subject classes the retrieval without restriction is nearly random before the first subject hit, and its FSHA is very close to unity. When the restriction is applied in the early T rounds of the query, the FSHAs increase to different degrees, depending on the class type, the accuracy of the classifier and the cutting round T. The small classes, *eyeglasses*, *mustache*, and *black* have more significant improvement than the large one, *female*. In addition, the improvement for all classes is very slight when the classification accuracies are lower than 80%. That is, we get the benefits from the approximated classification only with very

Table 1. Tested true semantic classes

class name	images	subjects a priori	
eyeglasses	262	126	15%
female	914	366	42%
mustache	256	81	9%
black	199	72	8%
whole database	2409	867	

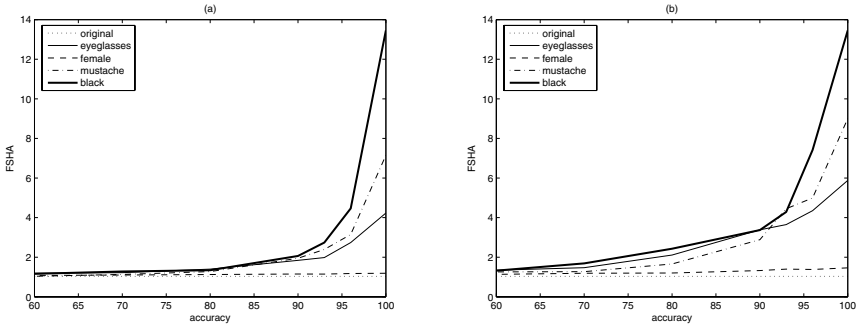

Fig. 1. FSHA with approximated classification of different accuracies. The predefined number of rounds before removing the restriction is $T = 10$ in (a) and $T = 20$ in (b)

accurate classifiers. This phenomenon becomes more evident when $T = 10$, i.e. the user's tolerance is smaller and the approximated classification is given up earlier and the whole database used instead.

3 Support Vector Machines

To obtain accurate classifiers especially for highly imbalanced data is not a trivial task. We adopt Support Vector Machines (SVMs) [8], which have shown good generalization performance in a number of diverse applications. In this section we give a brief introduction of SVM and describe how we have chosen its parameters.

3.1 Principles of SVM

Given a training set of instance-label pairs $(\mathbf{x}_i, y_i), i = 1, \ldots, l$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{1, -1\}$, the Support Vector Machines require the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{i:y_i=-1} \xi_i \\ \text{subject to } & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, i = 1, \ldots, l. \end{aligned} \tag{2}$$

Here the training vectors $\mathbf{x}_i$ are implicitly mapped into a higher dimensional space by the function ϕ. C_+ and C_- are positive penalty parameters of the error terms. The above problem is usually solved by introducing a set of Lagrange multipliers $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_l\}$:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \ & \mathbf{e}^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} \\ \text{subject to } & 0 \leq \alpha_i \leq C_+ \text{ if } y_i = 1, \\ & 0 \leq \alpha_i \leq C_- \text{ if } y_i = -1, \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0, \end{aligned} \tag{3}$$

where $\mathbf{e}$ is the vector of all ones, $\mathbf{Q}$ is an $l \times l$ positive semidefinite matrix given by $Q_{ij} \equiv y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$, where $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function. Then $\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i)$ and

$$\operatorname{sgn}(\mathbf{w}^T \phi(\mathbf{x}) + b) = \operatorname{sgn}\Big(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \Big) \tag{4}$$

is the decision function. For the kernel function $K(\cdot, \cdot)$, we have chosen the radial basis function (RBF) with common variance:

$$K_{RBF}(\mathbf{x}, \mathbf{z}; \gamma) = \exp\big(-\gamma \|\mathbf{x} - \mathbf{z}\|^2 \big) \tag{5}$$

because it has good classification power and only one parameter needs to be determined. We unify C_+ and C_- into a single parameter C with weights according to the inverse of their prior probability estimates, i.e. $C_+ = C$ and $C_- = C \cdot N^+ / N^-$, where N^+ and N^- are the numbers of the positive and negative labels, respectively.

3.2 Choosing SVM Parameters

The SVM experiments in this paper were implemented based on the *libsvm* library [1]. In the experiments it was noticed that the parameter settings have great impact on the performance of the resulting classifiers. There exist some parameter selection methods (e.g. [2]) which were reported to find good values for these parameters automatically, but it has also been shown that they are unstable in practice [3]. The gradient-based optimization algorithms adopted by these methods require a smoothing surface and a good starting point, which is albeit unknown beforehand. In addition, the penalty parameter C is incorporated into the kernel matrix, which is valid only when the SVMs are L_2 norm, but such SVMs for imbalanced data are not supported by most current SVM toolkits.

One can make use of some geomorphologic knowledge about the accuracy surface and then apply stochastic optimization to obtain a much more efficient parameter selection method. We first need a combined accuracy estimate which is proper for both a class and its complement. Given a true semantic class C_M and its complement $\bar{C}_M$, which are approximated by the restriction class C_R and $\bar{C}_R$, respectively, we adopt the minimum of the true positive accuracy and the true negative one, i.e.

$$\text{accu} = \min\Big(\frac{|C_R \cap C_M|}{|C_M|}, \frac{|\bar{C}_R \cap \bar{C}_M|}{|\bar{C}_M|} \Big). \tag{6}$$

Figure 2 illustrates an example of such accuracy measure's contour plot on the (C, γ)-plane. The example comes from 20-fold classification of the eyeglasses class using feature calculated from the left eye of each subject (see Section 4.1). The grid search used to draw this figure is not a part of the searching procedure, but helps us better understand the distribution of good values for C and γ.

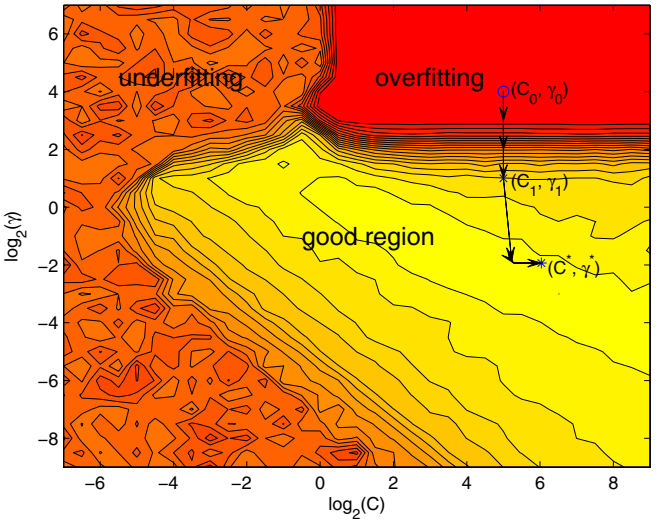

Fig. 2. Accuracy contour of eyeglasses in the (C, γ)-space with 20-fold cross-validation

Similar contour shapes have been also observed on a number of other real-world datasets (e.g. [1, 6]).

Based on the above understanding, we propose a path searching algorithm as follows: (1) Choose a large C and a large γ, i.e. a point on the overfitting plateau, for example, $C_0 = 2^5$ and $\gamma_0 = 2^4$. Then apply a line search downwards by decreasing γ until accu > 0.5. (2) Suppose the resulting point of step 1 is (C_1, γ_1), and for convenience, we write $\boldsymbol{\theta}_t = (C_t, \gamma_t)$. Denote $\mathbf{g}(t)$ the gradient of the accuracy surface and given $\Delta\boldsymbol{\theta}_1 = d_1\mathbf{g}(1)$, iteratively apply the conjugate gradient optimization procedure:

$$\Delta\boldsymbol{\theta}_t = \beta(t)\Delta\boldsymbol{\theta}_{t-1} + \mathbf{g}(t), \tag{7}$$

$$\beta(t) = \frac{\|\mathbf{g}^T(t)\|^2}{\|\mathbf{g}^T(t-1)\|^2}. \tag{8}$$

Step 1 locates a good starting point $\boldsymbol{\theta}_1$ for step 2. $\boldsymbol{\theta}_1$ is probably on the upper hill side of the good region mountain. The gradient $\mathbf{g}(t)$ at a point $\boldsymbol{\theta}_t$ is approximated by a one-sided finite difference where the change of accuracy is measured separately in the C and γ directions with difference magnitude h_k. A common form for the sequence h_k is $h_k = h/k^m$, where h and m are predefined positive constants. d_1 is the initial learning rate at (C_1, γ_1). If accu$(\boldsymbol{\theta}_t) >$ accu$(\boldsymbol{\theta}_{t-1})$ then record $\boldsymbol{\theta}_t$ and $\mathbf{g}(t)$, $t \leftarrow t+1$, $k \leftarrow k+1$. Otherwise, just shrink h_k by $k \leftarrow k+1$. This way we can obtain a path with only increasing accuracies. Note that the conjugate gradient method only works in low stochastic level. Therefore 20-fold cross-validation was used instead of the popular 5-fold setting.

The searching path for the eyeglasses class using the left eye feature is shown by arrows in Figure 2. Step 1 began with the initial point $(C_0, \gamma_0) = (32, 16)$

and the resulting point was $(C_1, \gamma_1) = (32, 2)$, from which the conjugate gradient optimization started with the setting $d_1 = 50$, $h = 1$, and $m = 0.1$. After 18 calls of the cross-validation procedure the searching algorithm returned the final point $(C^*, \gamma^*) = (65.7, 0.262)$ with accuracy 90.45%. The application of the procedure was the same for all other features and classes.

4 Experiments

The testbed we used in the experiments is our CBIR system named PicSOM [5], which utilizes the Self-Organizing Maps (SOMs) as the underlying indexing and relevance feedback processing technique. Some images are shown in each round of a query and the user is supposed to mark zero or more of them as relevant to the current retrieval task. The rest images in that round are treated as non-relevant. This relevance feedback is then used to form relevance score values in the best-matching map units (BMUs) corresponding to the shown images on each participating SOM. The effect of the hits is spread to the neighboring SOM units by low-pass filtering over the SOM surface.

More than one feature can be involved simultaneously and the PicSOM system has a separate trained SOM for each. The convolution provides implicit feature weighting because features that fail to coincide with the user's conceptions mix positive and negative user responses in the same or nearby map units. Such SOMs will consequently produce lower scores than those SOMs that match the user's expectations and impression of image similarity and thus produce areas or clusters of high positive response. The total scores for the candidate images are then obtained by simply summing up the mapwise values in their BMUs. Finally, a number of unseen images with the highest total scores are displayed to the user in the next round.

4.1 Data

In the FERET collection [7] the coordinates of the facial parts (eyes, nose and mouth) were obtained from the ground truth data, with which we calibrated the head rotation so that all faces were upright. All face boxes were normalized to the same size of 46×56 pixels, with fixed locations for left eye (31,24) and right eye (16,24) in accordance to the MPEG-7 standard [4]. The box sizes of the face and the facial parts are shown in the second column of Table 2.

After extracting the raw features within the boxes mentioned above, we applied Singular Value Decomposition (SVD) to obtain lower-dimensional eigenfeatures of the face and its parts. The numbers of the principle components preserved are shown in the third column of Table 2.

4.2 Single Classifier Results

The resulting 20-fold cross-validation accuracies and respective parameters for all tested classes using individual SVM classifiers are shown in Table 3. The

Table 2. Specification of the used features

feature name	normalized size	eigenfeature dimensions
face	46×56	150
left eye	24×16	30
right eye	24×16	30
nose	21×21	30
mouth	36×18	50

Table 3. True positive and true negative accuracies for individual classifiers with 20-fold cross-validation

class	face	left eye	right eye	nose	mouth
eyeglasses	77.10%, 88.22%	90.45%, 97.90%	90.84%, 97.62%	88.55%, 91.01%	—
	(181.02, 0.0030)	(65.67, 0.2617)	(63.29, 0.2196)	(0.54, 3.4822)	—
female	87.09%, 90.84%	82.17%, 82.81%	78.77%, 82.34%	68.93%, 71.04%	81.84%, 81.80%
	(1351.2, 0.0073)	(32.00, 0.1768)	(32.00, 0.1768)	(18.38, 0.1015)	(62.18, 0.1075)
mustache	78.52%, 78.17%	—	—	70.31%, 71.06%	84.38%, 87.46%
	(2.00, 0.0313)	—	—	(16.46, 0.1328)	(0.66, 0.0291)
black	79.90%, 85.48%	79.90%, 82.24%	77.89%, 79.91%	71.86%, 75.88%	80.40%, 84.71%
	(90.51, 0.0032)	(0.66, 2974)	(0.23, 0.6156)	(0.81, 0.0670)	(0.25, 0.1768)

first percentage in each cell is the accuracy for the true positive and the second for the true negative. The number pair under the accuracy percentages is the respective C and γ. It can be seen that the best accuracy for the eyeglasses class was obtained with the eye features, for the gender with the face feature, and for the mustache with the mouth feature. These results are consistent with our everyday experience. The case of the black race is not so obvious and all other features but the nose seem to perform equally well, but worse than for the three other classes.

4.3 Combining Individual Classifiers

Although the features used in the experiments are not fully uncorrelated, it is still beneficial to combine some of the individual classifiers to a stronger one. This can be done by performing majority voting weighted by their accuracies. For a specific class category, denote $L(f, I)$ the label to which an image I is classified by using the feature f and $\mathrm{accu}(f)$ the respective accuracy of that classifier. Assign j to I if

$$j = \underset{i}{\operatorname{argmax}} \Big\{ \sum_{L(f,I)=i} \big[\mathrm{accu}(f) - 0.5\big] \Big\}. \tag{9}$$

The subtractive term 0.5 is used here to give the best-performing classifiers extra reward compared to the worst-performing ones. Table 4 shows the accuracies after combination and the respective features used. It can be seen that for the classes of female and black the accuracies can be significantly improved

Table 4. Leave-one-subject-out true positive and true negative accuracies for combined classifiers

class	accuracy	features used
eyeglasses	95.91%, 96.88%	face, left eye, right eye, nose
female	90.62%, 94.58%	face, left eye, right eye, nose, mouth
mustache	84.11%, 87.78%	face, nose, mouth
black	85.70%, 91.04%	face, left eye, right eye, nose, mouth

by combining individual SVM classifiers. The combination also enhanced true positive accuracy for the eyeglasses class. By contrast, the accuracies of the mustache class after the combination remained at the same level as with the mouth feature only.

4.4 Obtainable FSHA Values

We obtained estimates of the FSHA with the combined classifiers by averaging the accuracies of the true positive and the true negative. This mean accuracy was then used when interpolating the FSHA values from the results of Section 2.2. The results shown in Table 5 indicate that the retrieval performance in terms of the first subject hit can be improved to different extent depending on the semantic criterion upon which the approximated classification is based. Also the number of rounds where the restriction is applied is a significant factor. The FSHA values for the eyeglasses class show clear improvement whereas the improvement for the female class alone is quite modest.

Table 5. FSHA estimates with the combined classifiers

	eyeglasses	female	mustache	black
T=10	4.2	1.2	1.6	2.0
T=20	5.9	1.3	2.3	3.4

5 Conclusions and Future Work

The possibility of incorporating auto-classification into interactive facial image retrieval was probed in this paper. We found that highly accurate classifiers are required to achieve significant advantage in terms of the first subject hit. Support Vector Machines were introduced into this task and we also proposed an automatic method to select good parameter values for training the SVM classifiers. The desired high accuracy can be achieved for a number of class categories by combining individual classifiers created with different low-level facial features. According to our results, we can speed up the occurrence of the first relevant hit by a factor up to nearly six in the case that the person we are searching for is wearing eyeglasses. With the other semantic classes tested, like

the black race or mustache, a bit lower level of improvement can be obtained. Note that even though the improvement by filtering the gender alone is not significant, it is in some cases possible to combine this highly accurate classifier with others to generate more specific semantic subclasses.

Due to limited data for the highly imbalanced classes, we had to use all frontal facial images in the FERET database for training the classifiers and their validation. The experiments to obtain the true values of FSHAs can be easily implemented after more independent external data is acquired, as we now have the operative classifiers available.

There is still plenty of space for further improvement. One of the major questions in the future will be how to handle the class categories with soft boundaries such as hairstyles. Furthermore, so far the class categories which satisfy the accuracy requirement are still limited because we only used five quite general low-level visual features. With the advance of feature extraction techniques we will obtain better classifiers and as a result, more semantic categories can be supported.

References

1. C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
2. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159, 2002.
3. C.-W. Hsu, C.-C. Chang, and C.-J. Lin. *A Practical Guide to Support Vector Classification.* Document available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
4. ISO/IEC. Information technology - Multimedia content description interface - Part 3: Visual. 15938-3:2002(E).
5. J. Laaksonen, M. Koskela, and E. Oja. PicSOM—self-organizing image retrieval with MPEG-7 content descriptors. *IEEE Transactions on Neural Network*, 13(4):841–853, 2002.
6. J.-H. Lee. Model selection of the bounded SVM formulation using the RBF kernel. Master's thesis, Department of Computer Science and Information Engineering, National Taiwan University, 2001.
7. P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J*, 16(5):295–306, 1998.
8. V. Vapnik. *Statistical Learning Theory.* NY: Wiley, New York, 1998.
9. Z. Yang and J. Laaksonen. Interactive retrieval in facial image database using Self-Organizing Maps. In *Proc. of IAPR Conference on Machine Vision Applications (MVA2005)*, Tsukuba Science City, Japan, May 2005.