Face Image Annotation Based on Latent Semantic Space and Rules

Hideaki Ito, Yuji Kawai, and Hiroyasu Koshimizu

School of Information Science and Technology, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
{itoh@sist, h104045@st, hiroyasu@sist}@chukyo-u.ac.jp

Abstract. This paper presents a face image annotation system based on latent semantic indexing and rules. To achieve annotation, visual and symbolic features are integrated. Two features are corresponding to lengths and/or widths of face parts and keywords, respectively. In order to develop annotation mechanism, it is required to vary the dimensions of the spaces which are constructed by the latent semantic indexing, and to represent direct relationships among features. Associated symbolic features to visual features are represented in rules based on decision trees. Co-occurrence relationships among keywords are represented in association rules.

Keywords: face image annotation, latent semantic indexing, decision tree, association rule.

1 Introduction

In recent years, image data are accumulated enormously, it is necessary to develop image annotation systems [2, 8]. Usually, objects or regions of natural images are described in words. However, in face image annotation systems, it is desired to represent inspired impressions from face images [4].

We have been developing a face image annotation system, named FIARS [6]. The purposes to develop this system are to retrieve face images in keywords, and to assign keywords to face images. Keywords are selected after emphasizing characters on faces. The characters are depicted by comparing one person and other persons. For example, when a caricature of a person is drawn, emphasis are made by appropriately modifying measured characters [1]. And, it is considered that the emphasized characters on the face are represented in terms of keywords, such as, round eye, thin lip, large nose, etc.

Annotation is achieved by integrating visual and symbolic features. Visual features are lengths and/or widths of individual face parts, called part data. Symbolic features are keywords which describe impressions with respect to sizes and/or shapes of a face. These features are integrated by latent semantic indexing [11] in FIARS. However, during progress of this system, it is required to specify associations from visual features to symbolic features in direct, and to adjust dimensions of constructed latent semantic spaces. For meeting them,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 766–773, 2008.

two mechanisms are developed; one is to construct rules, and another is to treat arbitrary dimensions of the spaces. Decision trees and association rules [5] are constructed. Decision trees specify direct relationships between part data and keywords. Association rules specify associations among keywords. On the other hand, retrieval results are changed according to the dimensions of the spaces.

Many annotation systems are developed by integrating two types of features. Rules to specify relationships between visual and symbolic features are proposed by [3]. Textures and colors are used as visual features. An annotation system based on latent semantic indexing is developed [13]. [9] utilizes probabilistic latent semantic indexing. On the other hand, many systems are developed for recognizing faces [1]. Some points or regions on a face are measured to meet inherent applications. In FIARS, lengths/widths and distances of face parts are measured, since characters are represented in keywords with respect to sizes or shapes. Moreover, an automatic facial expression analysis system is developed for human emotion analysis using face action units [10]. They are used to represent emotions, such as happiness, anger, fear, etc. FIARS deals with keywords which describe characters of face parts, such as thin lip, slender eyebrow, small eye, round face, etc.

This paper is organized as follows. Section 2 shows an overview of the system. Description of face images is shown in Sec. 3. Section 4 describes annotation mechanisms based on latent semantic indexing, decision trees, association rules, and their efficiencies. Finally, concluding remarks are described in Sec. 5.

2 An Overview

During the development of FIARS [6], it is desired to decide suitable dimensions of constructed spaces, and to describe relations among part data and keywords. The following three mechanisms are developed for meeting them. They are; to specify dimensions of constructed spaces; to make decision rules which specify relationships between part data and keywords, constructed from decision trees; and to make association rules which specify co-occurrence of keywords.

Figure 1 shows an overview of the system. When a face image is given, a set of keywords is assigned finally. To achieve this, a given face image is compared with existing faces. A set of face descriptions is stored in a face image database. Each description is specified in face images, part data and keywords. A collection of stored descriptions is used for constructing latent semantic spaces, decision trees and association rules.

The system constructs a numeric latent semantic space and a combined latent semantic space. A numeric latent semantic space is constructed from only part data, and a combined latent semantic space part data and keywords, respectively. A procedure for seeking keywords consists of the following steps. At first, some similar face images to a given face image are sought using the numeric latent semantic space. Next, a centroid of the obtained face images is computed in the combined latent semantic space, which is used as a query. Keywords similar to

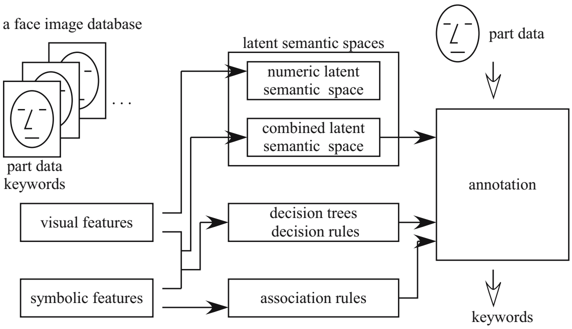


Fig. 1. An overview of annotation mechanisms and components of FIARS

this query are retrieved using the combined latent semantic space. Then, a cosine measurement is used. Retrieved keywords are seemed as keywords for the given face image.

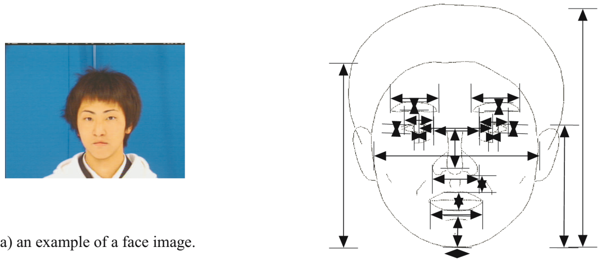
Decision rules represent whether a keyword is able to be assigned to a given face image, or not. On the other hand, association rules specify associations among keywords. When association rules are applied to a collection of keywords obtained by mechanisms of latent semantic spaces and decision trees, new keywords are captured by extending these keywords.

3 Face Description

An example of a face description is shown in Fig. 2. (a) shows an example of a face image. (b) presents part data. 24 places of the face parts are measured. (c) is a set of keywords which are assigned to (a). Keywords are restricted so that they represent sizes/shapes of face parts. Measurement places are set for measuring them. Moreover, when similar face images are retrieved using the numeric latent semantic space, another type of a feature, called point data, is considered [6]. The point data are given by distances between two points on the outline of a face. The number of measured places in the point data is greater than one in the part data. But, retrieval efficiencies of them are almost similar when each visual feature is used. Therefore, the part data are useful for their simplicity, and this feature is suitable to build decision trees.

Part data of face image I_d are represented in a vector, $\boldsymbol{v}_d = (v_{d,1}, \ldots, v_{d,24})^T$, called a part vector. From this vector, a normalized part vector is obtained, $\boldsymbol{v}'_d = (v'_{d,1}, \ldots, v'_{d,24})^T$. $v'_{d,j}$ is normalized value of $v_{d,j}$, and given by $v'_{d,j} = (v_{d,j} - \mu_j)/\sigma_j + 1/2, (j = 1, 24)$. μ_j and σ_j are the mean value and the standard derivation of face parts j, respectively.

On the other hand, 43 keywords are treated in current. For image I_d, keywords are represented in a keyword vector, $\boldsymbol{w}_d = (w_{d,1}, \ldots, w_{d,43})^T$. Each element $w_{d,j}$ is 1 or 0. They represent whether keyword j is defined in face image I_d, or not, respectively.



(a) an example of a face image.

(b) part data.

thin lip
slender eyebrow
small eye

(c) Examples of keywords which are assigned to face image (a).

length of the pupil of the left eye	1.2
length of the pupil of the right eye	1.3
length between the two pupils	6.0
length between the eyes	3.5
length of the left eye	3.2
. . .	
width of a face	13.2
length of a face	24.6
length of a face in visible	17.8
width of the chin	3.5

(d) part data, lengths/widths of 24 places of a face.

Fig. 2. An example of a face description

4 Annotation Mechanisms

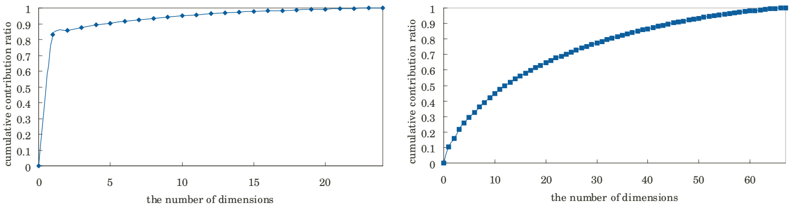
4.1 Dimensions of Latent Semantic Spaces

In latent semantic indexing, a matrix $A(m \times n)$ is decomposed into three matrices by the singular value decomposition [11], $A = U\Sigma V^T \cong U_k \Sigma_k V_k^T$. If the rank of A is r, then U is $m \times r$, the singular matrix Σ is $r \times r$, and V^T is $r \times n$. Let singular values be $\sigma_1, \ldots, \sigma_r$, and $\sigma_1 \geq \ldots \geq \sigma_r$. By selecting $k(1 \leq k \leq r)$, A is approximated to $U_k \Sigma_k V_k^T$. On the other hand, a cumulative contribution ratio is computed as $\Sigma_{j=1}^{k} \sigma_j / \Sigma_{j=1}^{r} \sigma_j$. k corresponds to the dimensions of a space. A contribution ratio seems to be useful because estimation of suitable dimensions is too hard [7].

To construct a numeric and a combined latent semantic spaces, matrix $\boldsymbol{N}$ and $\boldsymbol{C}$ are utilized, respectively. $\boldsymbol{N}$ is a collection of part vectors, $\boldsymbol{N} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d, \cdots, \boldsymbol{v}_n)$, where n is the number of stored face descriptions. $\boldsymbol{C}$ is a collection of the face description in terms of keywords and part data, $\boldsymbol{C} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_d, \cdots, \boldsymbol{c}_n)$. $\boldsymbol{c}_d$ is a concatenated vector $(\boldsymbol{w}_d; \boldsymbol{v}'_d)$, where $\boldsymbol{w}_d$ and $\boldsymbol{v}'_d$ are the keyword vector and the normalized part vector of I_d.

Figure 3 (a) and (b) show cumulative contribution ratios for the numeric latent semantic space and the combined latent semantic space, respectively. As shown in (a), it is seemed that individual part data are closely related each other. $\boldsymbol{N}$ can be reconstructed by a small number of dimensions in a sense that difference between the original matrix and the reconstructed matrix is little. As shown in (b), a large number of singular values are required for reconstructing $\boldsymbol{C}$ with little difference.

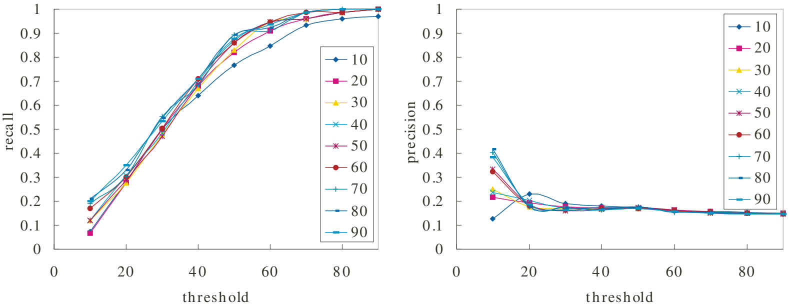
The dimension of each space is fixed in the developed system [6], which is equal to 3. Using these spaces we try to annotate new 10 face images. Figure 4 shows



(a) the cumulative contribution ratio in the numeric latent semantic space.

(b) the cumulative contribution ratio in the combined latent semantic space.

Fig. 3. Cumulative contribution ratios with respect to the numeric latent semantic space and the combined latent semantic space



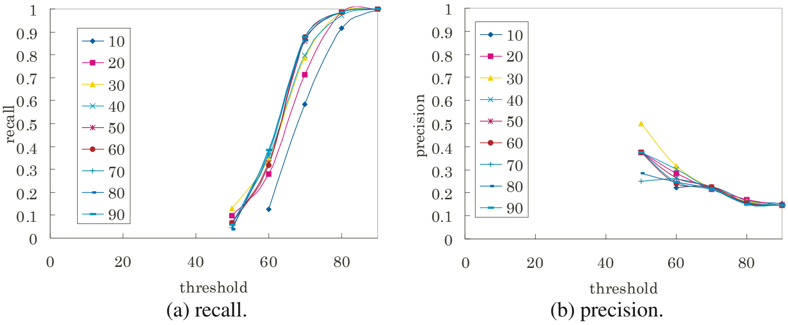
(a) recall.

(b) precision.

Fig. 4. Recall and precision when both dimensions of the numeric latent semantic space and the combined latent semantic space are equal to 3

recall and precision of retrieved keywords when thresholds are varied. (a) depicts recall, when a threshold for seeking keywords in the combined latent semantic space is changed from 10 to 90 degrees. The vertical axis and the horizontal axis are recall and a threshold for seeking keywords. At the same time, a threshold is changed for seeking similar face images in the numeric latent semantic space. Each line shows recall, when the thresholds for seeking similar face images are changed from 10 to 90 degrees. Moreover, (b) shows precision. When thresholds for keyword retrieval are increased, the recall are improved. The precision are low, although the thresholds are varied. When the thresholds for keyword retrieval are in during 30 and 90 degrees, precision are almost same although recall are improved. It seems that correct keywords are retrieved, but incorrect keywords are also retrieved as same as the correct ones.

Figure 5 shows recall and precision of retrieved keywords, when the dimensions of the numeric and the combined latent semantic space are equal to 3 and 33, respectively. The cumulative contribution of the combined latent semantic space is over 0.8 when the dimension is equal to 33. Thresholds for seeking faces and keywords are changed as described above. Under the thresholds less than 40 degrees for seeking keywords, keywords are not retrieved, since either similar face image or keyword is not retrieved. As shown in Fig. 4 and Fig. 5, the retrieval



(a) recall. (b) precision.

Fig. 5. Recall and precision when the dimension of the numeric latent semantic space is equal to 3 and the dimension of the combined latent semantic space is equal to 33

results shown in Fig. 5 are better than ones shown in Fig. 4. The recall shown in Fig. 5 are rapid increased, when the thresholds are increased. The precision shown in Fig. 5 are better than ones in Fig. 4, when the thresholds are around 50 degrees. To improve precision, it seems that the dimension of the combined latent semantic space is set high, and the threshold for seeking keywords is set low.

4.2 Decision Trees and Association Rules

Decision trees specify conditions to assign keywords in terms of part data. One decision tree consists of one root node, internal nodes, leaf nodes, and branches. The root node and the internal nodes are corresponding to individual part data, e.g., length of the pupil of the left eye, length between the two pupils, etc. Before a decision tree is constructed, simple discretization is applied to normalized part data. Their values are divided into three classes. About quarter of the stored face descriptions are assigned value 'a', half of them are assigned 'b', and the rest are assigned 'c' by ascending order. The values 'a', 'b' and 'c' are interpreted as 'small/short', 'middle' and 'large/long' depending on features.

During construction of a decision tree, a node is tried to be expanded using entropy, i.e., information gain [5]. If all face descriptions indicated by the leaf node are positive examples in a sense that a certain keyword is assigned to all of them, the leaf node is not expanded. Moreover, pruning is applied using an error ratio. An error ratio is computed as *(the number of negative examples at a leaf node) / (a total number of face descriptions at a leaf node)*. The negative examples are the face descriptions that the keyword is not assigned to. If an error ratio of a leaf node is less than a specified error ratio, the node is not expanded.

A decision rule is built from a decision tree directly, represented in $A \rightarrow B$. A is a set of patterns which are places of face parts. B is a keyword. If part data of a given face satisfy A then the keyword indicated in B is assigned to the face. For example, when an error ratio is 0 the following rule is captured;

```
height_of_the_left_eyebrow(b) and width_of_the_face(b) and
height_of_the_right_eyebrow(a) and
distance_between_the_jaw_and_the_line_of_centers_on_eyes(c)
-> slender_eyebrow
```

Table 1 shows recall and precision of captured keywords using decision rules, when an error ratio is changed. When an error ratio is small, the recall and the precision are almost constant. Although the recall is improved by increasing the error ratio, it is difficult to improve the precision.

Table 1. Recall and precision when an error ratio is changed

error ratio	recall	precision	error ratio	recall	precision
0.0	0.34	0.42	0.5	0.44	0.36
0.1	0.34	0.42	0.6	0.57	0.38
0.2	0.34	0.41	0.7	0.69	0.35
0.3	0.34	0.41	0.8	0.75	0.31
0.4	0.40	0.38	0.9	0.87	0.24

Table 2. Validity of captured association rules

support	confidence	the number rules	validity
0.01	0.65	121	0.84
0.03	0.45	122	0.71
0.05	0.20	124	0.51
0.06	0.15	102	0.53

On the other hand, an association rule $X \rightarrow Y$ represents co-occurrence relationships among keywords. X and Y are disjoint sets of keywords. To measure an association rule support and confidence are used [5]. For example, the following association rule is obtained;

```
small_eye and large_nose and dropping_eyes -> thin_lip
```

The support and the confidence of this rule are 0.04 and 0.79, respectively.

If the support is more than 0.01 and the confidence is more than 0.65, about 120 rules are obtained. Table 2 shows their validity. The validity of a rule is computed as *(the number of face descriptions including X and Y + the number of face descriptions including X and which are suitable to assign Y) / (the number of face descriptions including X)*. If confidence is high, validity is also high. Therefore, it is considered that suitable keywords are able be obtained by checking the confidences of rules.

5 Concluding Remarks

Three mechanisms for face image annotation are presented; they are latent semantic indexing, decision trees and association rules. When these techniques are applied to the same face image, it occurs that inconsistent keywords are obtained. For example, two keywords which have opposite meanings are obtained.

In current, these techniques are used independently, and these subsystems have been developing individually. An annotator uses these subsystems, and annotation results have to be integrated carefully.

We plan to integrate these mechanisms, as future works. Moreover, relationships among keywords are defined, such as synonyms, antonyms, broader terms and narrower terms, like a thesaurus. By specifying such relationships, assignment of inconsistent keywords will be prevented. Furthermore, to semi-automatically determine the values of some parameters for working subsystems are required, e.g., dimensions of spaces, an error ratio, etc.

Acknowledgement

We would like to thank Professor Iida, S. for helpful comments on this paper. The face images are used by permission of Softopia Japan, Research and Development Division, HOIP Laboratory. This research was supported in part by the Hibi Research Grant.

References

1. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and Machine Recognition of Faces: A Survey. Proceedings of the IEEE 83(5) (1995)
2. Datta, R., Ge, W., Li, J., Wang, Z.: Toward Bridging the Annotation-Retrieval Gap in Image Search. IEEE Multimedia (July-September, 2007)
3. Djeraba, C.: Association and Content-Based Retrieval. IEEE Tran. Knowledge and Data Engineering 15(1) (2003)
4. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. Pattern Recognition 36(3) (2003)
5. Han, J., Kamber, M.: Data Mining, Concepts and Techniques. Morgan Kaufmann, San Francisco (2006)
6. Ito, H., Koshimizu, H.: Some Experiments of Face Annotation Based on Latent Semantic Indexing in FIARS. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1208–1215. Springer, Heidelberg (2006)
7. Kontostathis, A., Pottenger, W.M.: A Framework for Understanding Latent Semantic Indexing (LSI) Performance. Inform. Processing & Management 42 (2006)
8. Li, J., Wang, J.Z.: Real-Time Computerized Annotation of Pictures. IEEE Tran. PAMI (to appear, 2007)
9. Monay, F., Gatica-Perez, D.: Modeling Semantic Aspects for Cross-Media Image Indexing. IEEE Tran. PAMI 29(10) (2007)
10. Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis From Static Face Images. IEEE Tran. SMC - Part B 34(3) (2004)
11. Skillicorn, D.: Understanding Complex Datasets. Data Mining with Matrix Decompositions. Chapman & Hall/CRC, Boca Raton (2007)
12. Softopia Japan Foundation: Face Image database, `http://www.hoip.jp/web=catalog/top.html`
13. Zhao, R., Grosky, W.I.: Narrowing the Semantic Gap? Improved Text-Based Web Document Retrieval Using Visual Features. IEEE Trans. on Multimedia 4(2) (2002)