Identification of Coreference Between Names and Faces

Koichi Yamada and **Kazunari Sugiyama**
Yasunori Yonamine and **Hiroshi Nakagawa**
Faculty of Engineering, Yokohama National University
79-5 Tokiwadai Hodogaya-ku Yokohama City, Kanagawa 240-8501 Japan
Phone: +81-45-339-4137
{aron, ksugi, yasunet, nakagawa}@naklab.dnj.ynu.ac.jp

Abstract

To retrieve multimedia contents by their meaning, it is necessary to use not only the contents of distinct media, such as image or language, but also a certain semantic relation holding between them. For this purpose, in this paper, we propose a method to find coreferences between human names in the article of newspaper and human faces in the accompanying photograph. The method we proposed is based on the machine learning and the hypothesis driven combining method for identifying names and corresponding faces. Our experimental results show that the recall and precision rate of our method are better than those of the system which uses information exclusively from either text media or image media.

1 Introduction

In multimedia contents retrieval, almost all of researches have focused on information extracted from single media, e.g. (Han and Myaeng, 1996) (Smeaton and Quigley, 1996). These methods don't take into account semantic relations, like coreference between faces and names, holding between the contents of individual media. In order to retrieve multimedia contents with this kind of relations, it is necessary to find out such relations.

In this research, we use photograph news articles distributed on the Internet(Mai, 1997) and develop a system which identifies a person's name in texts of this type of news articles and her/his face on the accompanying photograph image, based on 1) the machine learning technology applied to individual media contents to build decision trees which extract face regions and human names, and 2) hypothesis based combining method for the results extracted by decision trees of 1). Since, in general, the number of candidates from image and that from language are more than one, the output of our system is the coreference between a set of face regions and a set of names.

There are many researches in the area of human face recognition (Rowley et al., 1996)(Hunke, 1994)(Yang et al., 1997)(Turk and Pentland, 1991) and human name extraction, e.g. (MUC, 1995). However, almost all of them deal with the contents of single media and don't take into account the combination of multimedia contents. As a case of combining multimedia contents, there is a research of captioned images (Srihari and Burhans, 1994) (Srihari, 1995). Their system analyzes an image and the corresponding caption to identify the coreference between faces in the image and names in the caption. The text in their research is restricted to captions, which describes contents of the corresponding images. However, in newspapers or photo news, captions don't always exist and long captions like the captions used in their research are rare. Therefore, in general, we have to develop a method to capture effective linguistic expressions not from captions but from the body of text itself.

In the research field of the video contents retrieval, although there are many researches ((Flickner et al., 1995),etc), few researches have been done to combine image and language media (Satoh et al., 1997)(Satoh and Kanade, 1997)(Smith and Kanade, 1997)(Wactlar et al., 1996)(Smoliar and Zhang, 1994) . In this field, as language media, there are soundtracks or captions in the video or sometimes in its transcriptions. For analysis of video contents, the information which consists along the time axis is effective and is used in such systems. On the other hand, for analysis of still images, some other methods that are different from the methods for video contents retrieval are required because the relatively small amount of and limited information than information from videos are provided.

In section 2, the background and our system's overview are stated. In section 3 and 4, we describe the language module and the image module, respectively. Section 5 describes the com-

bining method of the results of the language module and the image module. In section 6, the experimental results are shown. Section 7 is our conclusions.

2 System architecture for combining

To find coreferences between names in the text and faces in the image of the same photograph news article, we have to extract human names from the text and recognize faces in the image (Figure 1).

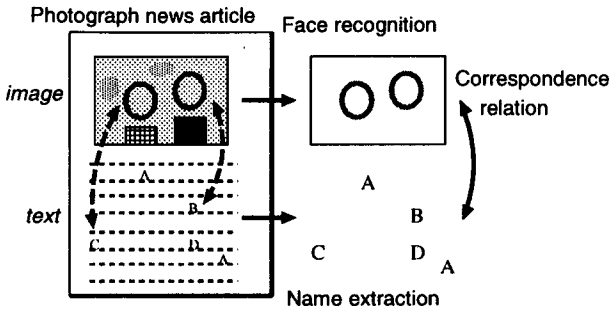

Figure 1: Human name extraction and face recognition.

The problem is that the face of the person whose name is appearing in a text is not always appearing in the image, and vice versa. Therefore, we have to develop a method by which we automatically extracts a person whose name appears in the text and simultaneously his/her face appears on the image of the same article. For the convenience, we define *common person*, *common name* and *common face* as follows.

Definition 1 *A person whose name and face appear in the text of the article and in the photo image of the same article respectively, is called* common person. *The name of the common person is called* common name, *and the face of the common person is called* common face.

This research is initiated by the intuition that is state as assumptions as follows:

Assumption 1 *The name of a common person has a certain linguistic feature in the text distinct from that of a non common person.*

Assumption 2 *The face of a common person has a certain image feature distinct from that of a non common person.*

These two assumptions are our starting point to seek out a method to identify the difference between the way of appearing of common names or faces in each media and the way of appearing of non common names or faces, and assign certainties of commonness to names and faces respectively based on the above assumptions.

Since each media requires its proper processing methodology, our system has the language module to process the text and the image module to process the image. Our system also has the combining module which derives the final certainty of a name and a face from the certainty of name calculated by the language module and the certainty of face calculated by the image module respectively.

For the image module, it is necessary to use the resulting information given by the language module, such as the number of names of high certainty, because the features of regions like where and how large they are, depend on the number of common persons. For example, the image module should select the largest region if the language module extracts only one name. On the other hand, for the language module, it is also necessary to use the result we get from the image module, such as the number of faces of high certainty, to select names of the common person.

However, if we consider the nature of these interactive procedures between the language module and the image module, it is easily known that one module cannot wait until the completion of analysis of the other module. To resolve this situation, we consider two kinds of method.

Method 1: First, the image (or language) module analyzes contents to proceed the process and outputs the partial results. Then assuming the result of the image (or language) module is correct, the language (or image) module analyzes the text (or image).

Needless to say, the assumed partial results might be wrong. In that case, the image (or language) module has to backtrack to resolve the conflict between the result of the image module and that of the language module. Namely, this method is a kind of search with backtrack and it also requires the threshold value by which the system decides whether the situation needs to backtrack or not. Moreover, the result depends on which media is analyzed first.

Method 2: Before combining of the results of image processing and those of language processing, the system works out all the hypotheses about the number of common

persons. Using all of these hypotheses, the system selects the best combination of the results. Its strong advantages are 1) the optimal solution is always found, and 2) each module can process independently.

Considering the advantages and the shortcomings of two of the above described methods, it is reasonable to adopt Method 2. In this research, the hypotheses of the number of common persons are "one", "two" and "more than two." The reasons of introducing "more than two" are the followings: the images containing four or more persons are very rare, and such images have similar features to the images containing three persons.

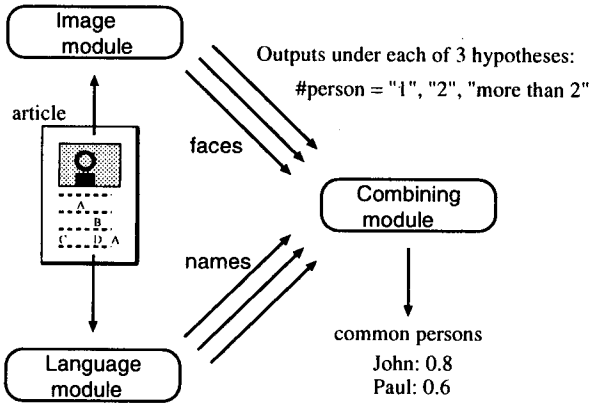

Figure 2: Overview of our system.

3 Extraction of human name candidates

The language module extracts the human name candidates from all human names appearing in the text and assigns certainty of commonness to each of the candidates of a common name. When the extracted name is a common name, the person is regarded as the important person in the article. Therefore, the linguistic expressions around the name probably have the specific linguistic features. Thus, our system decides whether an extracted name is a common name or not with information of the linguistic expressions around the human name. To select effective features for this purpose from the all features generated from the text, we employ a machine learning technique, because some important features could be fallen out if selected by hand. Moreover, machine learning technique might be able to learn incomprehensible phenomena for human.

It is hard to recognize meaningful linguistic features without morphological analysis. On the other hand, if the system does the syntax analysis, the handling of the ambiguity becomes a big problem. Furthermore, on the practical use, high processing cost becomes a problem to process huge amount of news articles. As the consequence, we adopt a word sequence pattern based approach. For this, firstly, we analyze texts of news articles with morphological analyzer JUMAN(version 3.6)(Kurohashi and Nagao, 1998) to extract the part of speech tags as the features in machine learning. Note that a compound noun is treated as one noun because if we treat component words of the compound noun individually, the patterns we have to deal with become too complicated for machine learning systems. The features to be used for learning are the followings.

Compound noun which contains a human name

The human name appearing in the news articles might have the adjacent words which describe additional information about the name such as title, age, year of birth and so on. The name with some kind of words, like title, sometimes becomes one compound noun and treated as one morpheme in our system. Our system tries to find this type of information as features for machine learning.

Part of speech tags around a human name

As well known, syntactic parsing is computationally heavy and usually has high ambiguities. Thus, instead of syntactic parsing, we extract the combination of a word, its part of speech tag and its relative position to the focused name for learning. Especially we focus on the words around the human name to capture the characteristic linguistic expressions about the human name. Our system employs two levels of the part of speech tag defined by the morphological analyzer JUMAN.

Since our system is for Japanese, object is described by a case particle. In pattern matching, instead of the sophisticated case analysis done by syntactic parsing, our system first applies the particle followed to the word as a feature. As for a predicate, we choose the predicate whose position is after the name and nearest to the name because in Japanese a predicate comes after subject, object, and other syntactic components.

Location and frequency of a human name

Location of the word is important because it reflects structures of documents. Our system uses features as follows: 1) whether the word is in the title or not, 2) the line number of the line the word is in, and 3) the number of the paragraph the word is in. Our system also uses the order of the occurrence of the name in all the name occurrences and the frequency of the name in the text.

Using linguistic features described above extracted from training data as inputs, we use C5.0 (Rul, 1998) to generate decision trees. For each case in test data, C5.0 outputs thè class predicted by the decision tree with the confidence of the prediction. We use the confidence as the output of this module.

Another factor for selecting feature for learning is how many morphemes around the name are used. In our experiment, ten morphemes around the name are used. The experimental results will be shown in section 6.

4 Extraction of human face candidates

To identify coreferences between the face in the image and the name in the text, this module should extract regions that are candidates of common face. In this section, we describe the image module which extracts face candidates from the image. The face candidates are the faces of persons who might be common persons. Next, as same as the language module does, this module learns the characteristic features of the region of a common face that are used to decide whether an extracted region as a face is a common face or not.

4.1 Extraction of face regions

To extract face regions, this module uses the following methods: 1) Filtering to remove noise, and 2) *RGB* based modeling of skin color to extract face region. Furthermore, this module generates features of each region and learns characteristics of the common face by C5.0. The value of each feature, e.g. location of face region, region size, depends upon the number of the persons appearing in the image as shown in Figure 3 and the text. To optimize feature based recognition, this module proceeds the processes corresponding to three hypotheses, say the number of common person is one, two, or more than two.

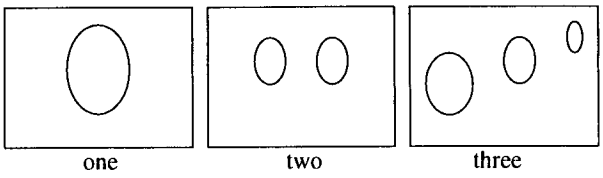

Figure 3: Differences in the features according to the number of the person.

4.2 Skin color modeling

The advantage of using color for face detection is robust against orientation, occlusion and intensities, and able to process fast, but the demerit is the difficulty in detecting only a face from a human body or other parts like hands, and to locate it accurately.

Darrell et al.(Darrell et al., 1998) convert (R,G,B) tuples into tuples of the form $(\log(G),\log(R) - \log(G),\log(B) - (\log(R) + \log(G))/2)$ which is called "log color-opponent space", and detect skin color by using a classifier with an empirically estimated Gaussian probability model of "skin" and "not-skin" in the space. Yang et al.(Yang and Waibel, 1995) develop a real-time face tracking system, and they propose an adaptive skin color model under different lighting condition based on the fact that its distribution under a certain lighting condition can be characterized by a multivariate Gaussian distribution(Yang et al., 1997). The variables are chromatic colors, that is, $r = R/(R+G+B)$, and $g = G/(R+G+B)$. On the other hand, Satoh et al.(Satoh et al., 1997) use the Gaussian distribution in (R,G,B) space in their face detection system because this model is more sensitive to brightness of skin color.

The picture of the newspaper we treat is a scene picture that includes not only a common face but also other faces, and a face doesn't always look straight forward. Thus, we use color information to detect a face because the color doesn't depend on its orientation. Suppose that the skin color distribution complies with the Gaussian distribution in (R,G,B) space(Satoh et al., 1997). Then, we introduce the Mahalanobis distance. That is the distance from the center of gravity of the group considering variance-covariance of data. We calculate the mean intensity $M(=(\hat{R},\hat{G},\hat{B})^T)$, variance-covariance matrix V and Mahalanobis distance d from skin color data of $5pixel \times 5pixel$ blocks, which are extracted from the cheek colored areas of 85 persons (Satoh et al., 1997). The almost all of cheek colored areas express natural

skin color and they are rarely in a shadow even if the people wear hat, etc. Suppose $\boldsymbol{I}$ be intensities of a pixel of the input image. Then, if that pixel satisfies (1), we take that pixel as the candidate pixel with skin color.

$$d^2 > (\boldsymbol{I} - \boldsymbol{M})^T \boldsymbol{V}^{-1} (\boldsymbol{I} - \boldsymbol{M}) \qquad (1)$$

where value d is experimentally optimized.

The method we described above is not so accurate in some cases. Some extra, non-facial regions would also be extracted simultaneously. To achieve higher accuracies, we examine the distribution of $(R+G+B)-(R-B)$, and draw border lines in order to contain more than 80% of the sample. We decide the triangle manually by observing the various output images. We extract the pixels which is in the triangle shown in Figure 4.

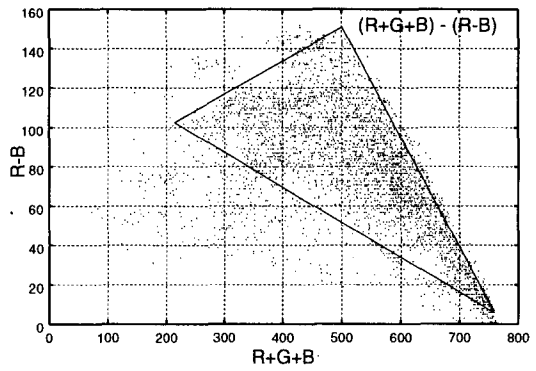

Figure 4: Skin color in $(R+G+B)-(R-B)$ space.

Some results by this method is shown in Figure 5. As you can see, not only faces but also hands and other regions whose color is similar to skin color are also extracted. To eliminate these undesirable regions, we use a decision trees built by C5.0 as stated in 4.3.

Figure 5: Result of face candidate region extraction.

4.3 Features of extracted regions

In this research, we use the following 17 features including the composition information of the whole image, in addition to the form and color of the region that is used with conventional image retrieval(Han and Myaeng, 1996). The following five features are used to express the form of skin color region: 1) Ratio of region to the largest region, 2) Ratio between the length of X-axis direction and the length of Y-axis direction, 3) Rectangularity, 4) Ellipticity, 5) Eccentricity.

The feature about the color is the followings:

6-9) Each of the mean of R, G, B and intensity Y.

The following eight features are positional information of the region.

10) Aspect ratio of the whole image.

11,12) x, y coordinates of the center of gravity of the region.

13) Distance between the center of gravity of the region and the center of the whole image, normalized with a half of the length of the diagonal line of the image.

14) The order of the region in descending order of 13).

15) Distance between the center of gravity of the region and the center of the upper end of the whole image, normalized with the length from the center of the upper end to the left lower end (or the right lower end).

16) The order of the region in descending order of 15).

17) Suppose that the image is divided into 3×3 sub-areas. Which of these sub-areas the center of gravity is in.

Using these 17 features extracted from training data as input features, we use C5.0 to learn decision trees, which extract candidates of common face with different certainties as described in section 3. The experimental results will be shown in section 6.

5 Combining candidates from image and language

In this section, we describe the combining module whose inputs are the candidates extracted by the language module and the image module described in section 3 and 4, respectively. Its output is the result of the whole system.

5.1 Input

As already said, since the language module and the image module process under hypothesis of "one", "two", or "more than two" persons, respectively, one module outputs three results according to these three hypotheses. Then outputs from both modules are expressed as follows:

$$\textit{(output of language module)} = f_{lang}(n, x) \quad (2)$$

$$\textit{(output of image module)} = f_{image}(m, y) \quad (3)$$

Note that n and m are the number of the common persons adopted as the hypothesis. x and y are orders in ascending order of certainty about the person being common. The certainty of the decision in the language module and the image module is the confidence output by C5.0. For example, $f_{lang}(n, 2)$ expresses the certainty of the person who has the second highest certainty. Each output is something like the graph on Figure 6. In this figure, all of the extracted

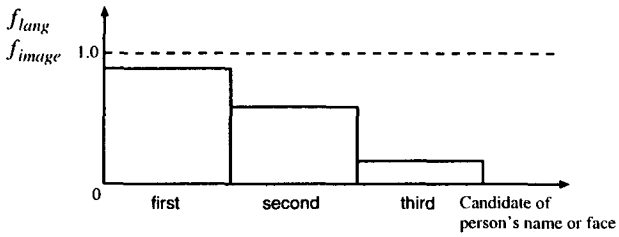

Figure 6: Output from each module under one hypothesis.

candidate names or faces are sorted in descending order of calculated certainties by distinct decision trees of the language module or the image module because the number of the common person might be more than one. By introducing certainties, as later described, we obtain enormous flexibility in combining candidates from the language module and those from the image module.

5.2 Combination of hypotheses

Since the language module and the image module process under each of three hypotheses, there are 3×3 combinations of the results. This combining module selects the best pair from those combinations and outputs the results based on the selected pair. To select the best pair, we introduce some kinds of distance described as follows.

Distance between outputs of two media

The distance between the result of the image module and the result of the language module f_{li} is defined by (4).

$$f_{li}(n, m) = \sum_{z=1}^{M} \frac{|f_{lang}(n, z) - f_{image}(m, z)|}{f_{lang}(n, z) + f_{image}(m, z)} \quad (4)$$

where M is the maximum number of the persons known from the results of both modules. As you know from (4), the nearer the certainties of the candidates from the language module and the image module which have the same order z are, the smaller the $f_{li}(n, m)$ is.

Distance between output of media and hypothesis

If there is difference between a hypothesis and the output calculated under the hypothesis, say f_{lang} and f_{image}, the hypothesis should not be considered to be valid. Therefore, we introduce the distance between the hypothesis and the output of the language module: f_{lang} or that of the image module: f_{image}. A hypothesis of n common persons is defined in (5).

$$f_a(n, x) = \begin{cases} 1 & (x \leq n) \\ 0 & (x > n) \end{cases} \quad (5)$$

where x is the order of certainty of candidates. Since each of the language module and the image module has its own hypothesis, the combining module calculates the distance f_{al} defined by (6) between the hypothesis used in the language module and the result from the language module. It also calculates the distance f_{ai} defined by (7) between the hypothesis used in the image module and the result from the image module.

$$f_{al}(n) = \sum_{z=1}^{3} \frac{|f_{lang}(n, z) - f_a(n, z)|}{f_{lang}(n, z) + f_a(n, z)} \quad (6)$$

$$f_{ai}(m) = \sum_{z=1}^{3} \frac{|f_{image}(m, z) - f_a(m, z)|}{f_{image}(m, z) + f_a(m, z)} \quad (7)$$

In the case that the hypothesis is "more than two", the certainty of candidates whose order is fourth or larger are ignored.

Decreasing factor for each inconsistent hypothesis

Different hypotheses of the language module and the image module indicate inconsistency. However, since the analysis of each module is not perfect, our system does not exclude such

inconsistent combinations of hypotheses. Instead, we decrease the certainty of such inconsistent combinations. For this, we use decreasing factors $D(m,n)$ where n and m mean the hypothesized number of person in the language module and the image module, respectively. We empirically tuned up the actual values of $D(m,n)$ as shown in Table 1.

Table 1: Decreasing factor $D(m,n)$ for each inconsistent hypothesis.

		n		
		1	2	3 or more
m	1	1.0	0.9	0.5
	2	0.9	1.0	0.6
	3 or more	0.5	0.6	0.8

Integration of the measures

Using these three distances, namely f_{li}, f_{al} and f_{ai}, and $D(n,m)$, the combining module finally calculates total certainty $f(n,m)$ defined by (8) for each combination of hypotheses. The smaller the $f(n,m)$ is, the nearer the result from the language module is the result of the image module.

$$f(n,m) = \{f_{li}(n,m)+1\}\{f_{al}(n)+1\}\{f_{ai}(m)+1\} \times \frac{1}{D(n,m)} \quad (8)$$

5.3 Combining the results

When a combination which has the smallest $f(n,m)$ has been selected, the results from the language module and the image module are fixed. The system combines these results into one result $f_{union}(n,m,z)$, where the person corresponding to z is expected to be a common person. $f_{union}(n,m,z)$ is the final output of the whole system. For this combining, we investigate two methods as follows. In (9), the consistency on the number of common persons is regarded as an important factor. On the other hand, in (10), when at least one of two module, namely the language module or the image module, assigns high certainty to a candidate person, the whole system finally assigns high certainty to the candidate person.

$$\forall z, f_{union}(n,m,z) = f_{lang}(n,z) \times f_{image}(m,z) \quad (9)$$

$$\forall z, f_{union}(n,m,z) = 1 - \{1 - f_{lang}(n,z)\}\{1 - f_{image}(m,z)\} \quad (10)$$

The final outputs of whole system are something like these: "John: common person (certainty: 0.8)", "Paul: common person (certainty: 0.4)" and so on. These results are used to find the face on the image if we specify a certain name in the text to retrieve his/her face image, or vice versa.

6 Experiments

We have experimentally evaluated the system we proposed by comparing with the simple systems which contain only the language module or the image module respectively to confirm the effect of the combining process. The language module and the image module work under three kinds of hypothesis in the simple systems as well. Thus, we use the system's result which has the minimum distance between the output of media and the hypothesis defined by formula (6),(7) as the baseline of evaluation. In our experiments, we use the photograph news in the web page called "AULOS" distributed by The Mainichi Newspapers(Mai, 1997). The average length of the text of the article is about 300 characters or 100 words. The almost all of the images are full colored, and the average size of them is about 250 × 200 pixels. Moreover, the images are not accompanied with captions. On this evaluation, we use articles with full colored images published on May and June 1997. As for common name extraction, we did four fold cross-validation for 228 articles of this period which contains common human names. As for common face extraction, we did three fold cross-validation for the set of color photograph images which are contained by the articles used by the language module. To evaluate how accurate the system identifies the given person being a common person, we calculated the recall and precision rate of the system's decision about a person being common. Since the outputs of our system are certainties, recall and precision rates are defined as follows.

$$Recall = \frac{\sum_{i \in cc} W(i)}{Number\ of\ the\ common\ persons} \quad (11)$$

$$Precision = \frac{\sum_{i \in cc} W(i)}{\sum_{\forall i} W(i)} \quad (12)$$

where $W(i)$ is the certainty of person i, and cc means a set of all correctly identified persons.

Table 2: The evaluation results of the outputs from each module.

	Recall	Precision
Language module	0.68	0.67
Image module	0.52	0.64
Combining module based on (9)	0.42	0.74
Combining module based on (10)	0.76	0.69

The evaluation results of each module is shown in Table 2.

For the language module and the combining module, we evaluate names and its certainties. On the other hand, for the image module, we evaluate only certainties under the assumption that the human name of the face which was assigned higher certainty is correct because the image module doesn't output human names. The effect of combining appears as the difference between the results of the combining module and the results of the language module or the image module. The combining module has two variations. The module based on (10) improved both recall and precision rates by combining. The reason of high recall rate is that one module picks up the person whom the other module fails to pick up. Since high precision rate is maintained, this compensation is really effective. On the other hand, the combining module based on (9) improves the precision rate more than the module based on (10). The reason of this phenomena is that the module is able to cancel the noise which appears in one media contents by the other media contents. However, the recall rate was decreased as expected from (9).

7 Conclusions

We have developed the system which identifies coreferences between the human face in the image and the human name in the text by selecting combinations of hypotheses and the combining of the results from the language module and the image module. The experimental result is that recall is 42% to 76% and precision is 69% to 74%. This result indicates that the practical use of semi-automatic extraction of common person from multimedia contents for IR purposes would come into our sight with some technical improvement along this line of research strategy.

References

T. Darrell, G. Gordon, M. Harville, and J. Woodfill. 1998. Integrated person tracking using stereo, color, and pattern detection. *CVPR'98*, pages 601–609.

Myron Flickner, Harpreet Sawhney, et al. 1995. Query by image and video content: The QBIC system. *Computer*, 28(9):23–32.

Kyung-Ah Han and Sung-Hyun Myaeng. 1996. Image organization and retrieval with automatically constructed feature vectors. *SIGIR'96*, pages 157–165.

H. M. Hunke. 1994. Locating and tracking of human faces with neural networks. Tech. Report CMU-CS-94-155, Carnegie Mellon University.

Sadao Kurohashi and Makoto Nagao. 1998. Japanese morphological analysis system JUMAN manual (version 3.6). Kyoto University.

The Mainichi Newspapers, 1997. *AULOS Photo News.* http://www.mainichi.co.jp/.

DARPA, 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6).*

H. A. Rowley, S. Baluja, and Takeo Kanade. 1996. Neural network-based face detection. *CVPR'96*, pages 203–208.

RuleQuest Research Pty Ltd, 1998. *See5 / C5.0.* http://www.rulequest.com/.

Shin'ichi Satoh and Takeo Kanade. 1997. Name-It: Association of face and name in video. *CVPR'97*, pages 368–373.

Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade. 1997. Name-It: Naming and detecting faces in video by the integration of image and natural language processing. *IJCAI-97*, pages 1488–1493.

Alan F. Smeaton and Ian Quigley. 1996. Experiments on using semantic distances between words in image caption retrieval. *SIGIR'96*, pages 174–180.

Michael A. Smith and Takeo Kanade. 1997. Video skimming and characterization through the combination of image and language understanding techniques. *CVPR'97*, pages 775–781.

Stephen W. Smoliar and HongJiang Zhang. 1994. Content-based video indexing and retrieval. *Multimedia*, 1(2):62–72.

Rohini K. Srihari and Debra T. Burhans. 1994. Visual semantics: Extracting visual information from text accompanying pictures. *AAAI-94*, 1:793–798.

Rohini K. Srihari. 1995. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56.

M. Turk and A. Pentland. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Howard D. Wactlar, Takeo Kanade, Michael A. Smith, and Scott M. Stevens. 1996. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52.

Jie Yang and Alex Waibel. 1995. Tracking human faces in real-time. Tech. Report CMU-CS-95-210, Carnegie Mellon University.

Jie Yang, Weier Lu, and Alex Waibel. 1997. Skin-color modeling and adaptation. Tech. Report CMU-CS-97-146, Carnegie Mellon University.