In: Proceedings of the IEEE International Workshop on Robot-Human Interactive Communication, pages 436-441, Bordeaux/Paris, 2001.

An Adaptive Person Recognition System

Kazunori Okada[†] & Lawrence Kite[†] & Christoph von der Malsburg[†,§]
[†] Computer Science Department, University of Southern California, USA
[§] Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
{kazunori,lkite,malsburg}@selforg.usc.edu

Abstract

We propose a framework for integrating the processes of object recognition and knowledge adaptation. This framework acknowledges that the performance of the object recognition process depends directly on the state of the system's internal knowledge, i.e., its memory, and conversely, that the efficacy of the system's knowledge adaptation process is enhanced by its ability to recognize objects with greater accuracy. Thus, object recognition and knowledge adaptation are inseparable aspects of the same cognitive task, and must be coordinated if the system is to be robust in the context of objects that present variations in illumination, scale, shift, and other temporal variations. Specifically, the presented system combines a multiple-cue person recognition component and an example-based knowledge adaptation component, and is applied to the task of automatic video indexing of personal appearance events, i.e., the presence of a full frontal face in a video stream. We present the details of this integrated framework and demonstrate successful experimental results.

1 Introduction

The task of object recognition is to associate sensory inputs to some internal knowledge about known objects stored in memory. This association provides us with information required to interact with the environment. Autonomous and adaptive knowledge acquisition from raw sensory data is another important aspect of the visual system. Newly encountered objects must be added to previously acquired internal knowledge. Furthermore, since the appearance of objects may change continuously, the knowledge of such objects has to be incrementally updated to maintain accurate representations. Our investigation focuses on the fact that the processes of object recognition and knowledge adaptation are not independent. The state of the internal knowledge constrains the performance of the recognition process. In turn, the results of the recognition process provide a basis for the knowledge adaptation process.

In this paper, we investigate methods for realizing the interdependent structure in a classical example-based pattern classification system. Such a system integrates the recognition and adaptation processes, introducing a dynamic relation between recognition performance and the state of internal knowledge. In this framework, a system learns while performing; the online incremental knowledge adaptation facilitates robust recognition of objects that undergo intrinsic temporal variations. In pattern recognition, the example-based approach has been one of the most common recognition architectures. It represents a known object by a set of examples or templates and performs a nearest neighbor search for identifying an arbitrary input with the known object most similar to the input. Most previous studies in this domain, however, have treated the knowledge as a *static* and *manually constructed* database [1]. Thus the performance of these systems depends on the specific choice of database. Weng and Hwang [11] proposed an incremental learning system of a facial knowledge database, which employs Linear Discriminant Analysis for recursively partitioning the input feature space, which is given by Principal Component Analysis. These statistical algorithms are usually time-consuming and require recomputation of the internal knowledge each time a new sample is added. An example-based approach simplifies the implementation of incremental learning since the previously acquired knowledge can be modified simply by the addition or subtraction of templates.

We applied the proposed method to an automatic video indexing problem. A video indexing system takes a video stream as input and extracts event, which serve as symbolic indices of a visual database, thereby reducing the search-time of the database. Be-

cause an input video stream may contain a wide variety of image variations, automation of this task is one of the most challenging problems in computer vision. In general, the definition of events includes many objects and their behavioral states [8]. In this study, we concentrate on events of *personal appearance*, which provide information of *who* appears *when* [3]. Our system utilizes *spatiotemporal* segmentation for extracting the events. Each segmented person is then recognized with an example-based *known-person database* that is continuously adapted by the results of this recognition process. Sato and Kanade [7] demonstrated a technique for indexing facial identities by associating co-occurrence of faces in a visual stream and names in corresponding closed-captions. They did not address, however, the issue of incremental knowledge acquisition based on visual information.

2 The System

The proposed system consists of two stages: spatiotemporal segmentation and integration of object recognition and knowledge adaptation. The former detects personal appearance events from a continuous video stream. The latter provides identity information for the detected events and adapts incrementally a known-person database.

2.1 Spatiotemporal Segmentation

Facial regions within each frame are detected by a coarse-to-fine search using motion, convex shape and facial similarity cues [9]. First, the motion and convex shape cues are applied to each frame, resulting in a set of region of interests (ROIs). Each ROI is cropped and normalized to a fixed size (128×128 pixels). Next, bunch graph matching [12] is performed on these normalized ROIs with a coarse graph (16 nodes). The similarity of the bunch graph to each ROI is used as a confidence measure for the presence of a frontal face. A threshold function is applied to this confidence value, determining whether a given ROI contains a face.

Space and time discontinuity cues of the face trajectory, which tracks smoothly moving faces in consecutive input frames, are used for segmenting the continuous input video stream. The trajectory is created when a new face is found and is discontinued when either 1) no face is found within a current frame (time discontinuity) or 2) a spatial displacement of a face between two consecutive frames exceeds a proximity threshold (space discontinuity). This spatiotemporal segmentation results in a set of sub-sequences, each of which contains the face of only *one* person. A personal appearance event is defined by the detected subsequence. See Okada and von der Malsburg [3] for more details.

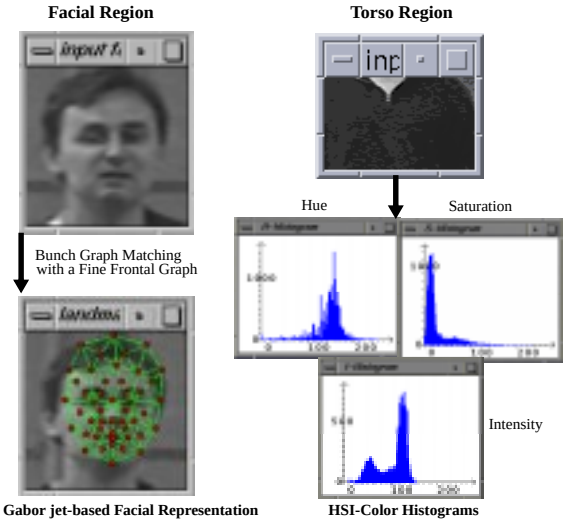

Figure 1: A view representation of an image containing a person.

2.2 Representation of Persons

Each detected sub-sequence is represented by two formats: 1) a sequence of Gabor-jet based facial representations [12] and 2) a sequence of color histograms of the corresponding torso regions. Each frame of the former sequence is subjected to a bunch graph matching with a fine graph (48 nodes). The torso region is cropped from the original color frame at a location determined heuristically from the detected facial position. For each of such frames, histograms of color pixel values are computed separately for each component of the Hue-Saturation-Intensity (HSI) color space. We call the representation of a single video frame by the pair of the facial representation and the torso-color histograms, the *view representation*, as illustrated in figure 1.

2.3 Integration of Object Recognition and Knowledge Adaptation

The second stage of the system performs three processes: 1) *recognition*, 2) *knowledge adaptation*, and 3) *forgetting*. Each process is based on the interaction between a detected sub-sequence and the known-person database. The known-person database contains a set of entries, each of which represents a single person. Each such entry consists of a set of view representations and a prototype. The prototype is defined as the average of all the view representations in the entry.

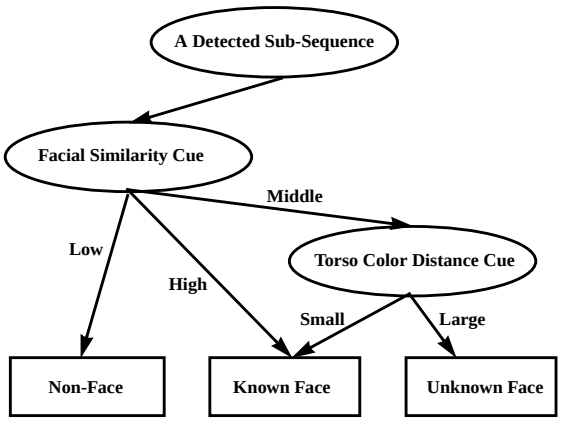

Figure 2: An illustration of the recognition process with our example-based multi-cue system.

2.3.1 Recognition Process

The recognition process classifies a detected sub-sequence into one of three cases: 1) *a known person*, 2) *an unknown person*, or 3) *a non-person*. In the known-person case, the sub-sequence is identified as one of the persons already stored in the known-person database. In the unknown-person case, the sub-sequence is determined to be a person not previously stored in the database. The non-person case means that the sub-sequence was erroneously detected and did not contain a person. This classification is carried out by sequentially applying threshold functions to multi-cue similarities between the sub-sequence and the database entries. The sub-sequence is first analyzed by the facial similarity cue. The torso-color distance cue is used only when the facial similarity cue cannot provide sufficient information. Figure 2 illustrates this classification process.

For the facial similarity analysis, a frame-to-frame similarity value between a frame in the sub-sequence, and a facial representation in an entry of the known-person database is computed by averaging the normalized dot-products of two corresponding local feature vectors (jets) over every node of the facial graph. For each frame, the maximum of similarity values over all the facial representations in a database entry is taken as a frame-to-entry similarity value. The sequence-to-entry similarity for each entry is given by averaging the maximum similarities over all frames of the sub-sequence. The database entry which gives the highest sequence-to-entry similarity value is considered as an identity candidate and the similarity value is considered as its confidence value. This identity confidence is subjected to a decision function with two thresholds. If the confidence exceeds the higher threshold, the detected sub-sequence is identified as the computed candidate. If the confidence fails to reach the lower threshold, the sub-sequence is classified as a non-person. If the confidence falls between the two thresholds, the decision is deferred to the torso-color analysis.

For the torso-color analysis, a sequence-to-entry distance value of the sub-sequence to the identity candidate entry is computed in the same manner as for the facial similarity analysis. The frame-to-frame distance is given by averaging the Kolmogorov distances [10] of the Hue and Saturation histograms. In order to mitigate influences from illumination variations, the Intensity histogram is not used. The average torso-color distance is subjected to another threshold function, wherein the input is identified as the candidate if the input torso-color is very similar to that of the candidate. Otherwise, the input is recognized as a previously unknown person. The threshold for the torso-color distance is set with a qualitatively more strict criterion than for the facial similarity, in order to reduce the risk of false identifications caused only by similar torso-color. The identity of the detected sub-sequence is labeled according to the classification result described above.

2.3.2 Knowledge Adaptation Process

For each of the three classification cases, a different adaptation process is performed. Following terminology of Piaget's theory of infant sensorimotor development, we call the adaptation of a known person, *assimilation*, and of an unknown person, *accommodation* [6]. The known-person database is adapted in each of these cases. When the sub-sequence is classified as a non-person, however, the database is not modified. The assimilation process updates the database entry for the known person identified by the recognition process. It adds the sub-sequence to the known person's view representations and updates the known person's prototype by computing a new average of the view representations. The accommodation process adds the sub-sequence to the known-person database as a new entry. The new entry consists of all view representations of the sub-sequence and is associated with a new identity label. When the system starts to process a new video stream, the known-person database is initialized to include no entries. Therefore, the first detected sub-sequence cannot be identified and is accommodated automatically. The rest of the recognized sub-sequences will be accommodated or assimilated according to the above-described rules.

2.3.3 Forgetting Process

Note that the knowledge adaptation process outlined above only can increase the size of the known-person database as time goes on. In order to process a very long, or even an infinitely long, video input within a limited memory resource, our system must be able to discard, or *forget*, some information that becomes redundant or irrelevant over time. The forgetting process is a part of the knowledge adaptation process which counters the divergence of the size of the database. Each time a detected sub-sequence is assimilated or accommodated, the number of view representations of each database entry is decreased by a certain number and its prototype is re-computed.

For this reduction process, the view representations of each entry are first sorted according to their similarity to the entry's prototype. The process discards view representations that give similarity values that are lower than others or that have larger deviations from the mean. The number of entries to be discarded is determined by the following function,

$$\begin{aligned} D_i = r_i N_i &= f(t_i) g(M, N_i) N_i, \\ f(t_i) &= \tfrac{t_i}{t_i + a}, \\ g(M, N_i) &= \tfrac{1}{1 + \exp(-bMN_i - c)}, \end{aligned} \tag{1}$$

where D_i denotes the number of views to be discarded in the i-th entry; N_i denotes the current number of views in the entry; r_i, ranging from 0 to 1, is the forgetting rate of the entry; M denotes the current number of known-persons; t_i denotes the temporal distance, which is defined as the number of persons seen by the system since the i-th entry was last recognized; a, b and c denote free parameters.

Figure 3 plots the functions $f(t_i)$ and $g(M, N_i)$. The former realizes the effect of temporal memory decay. More views will be forgotten when the entry was last visited longer ago. A queue is used to keep track of when each known-person was last assimilated or accommodated. The parameter a controls the decay rate. On the other hand, the latter implements the effect of data size convergence. It specifically discards more views from an entry when the entry includes more views and the database includes more known persons. Moreover, before the database size reaches the parameter c, the database size decreases slowly; it decreases faster when the size exceeds c. The parameter b controls the slope of this transition. The forgetting rate is determined by modulating these two functions with each other. When a known-person entry loses all of its view representations, it is removed from the database and is forgotten completely.

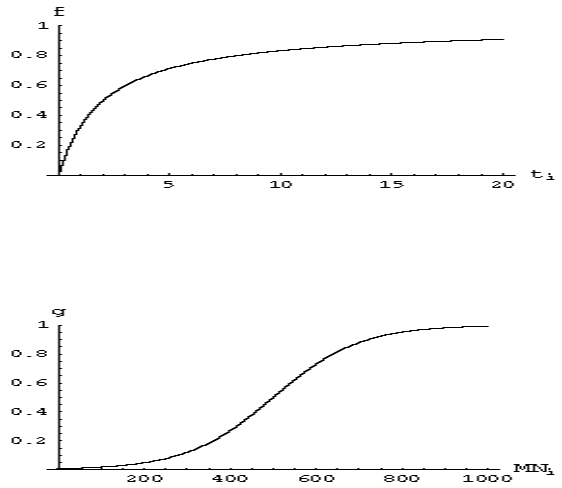

Figure 3: Two components of the forgetting function. The top plot displays the temporal memory decay function $f(t_i)$, and the bottom plot displays the data size convergence function $g(M, N_i)$. The free parameters, a, b and c, are set to 2.0, 0.01 and 500, respectively.

3 Experiments

Our proposed system has been tested with scenes from a podium speech setting with freely moving speakers. In such a setting, the movement of speakers creates a variety of image variations (e.g. translation, scaling, depth rotation, and expression). Moreover, speakers may sometimes move in and out of the field of view, and the illumination condition may change drastically.

Figure 4 illustrates an example of our system's output. A video stream was recorded by a static camcorder in a seminar room environment. It consists of 1900 frames and contains three speakers, speaking at a podium, moving freely, and reappearing to the podium occasionally. Our system successfully extracted seven personal appearance events from this input, identified each person correctly, and automatically adapted the correct known-person entry in the database. Consecutive events identified as the same person are merged into a single event. Frame gaps between the events corresponds to situations where either no speaker was present in the stream or the rotation of a speaker's head was very large. The average throughput of this system was approximately two frames per second on an Intel PC with a 733MHz processor.

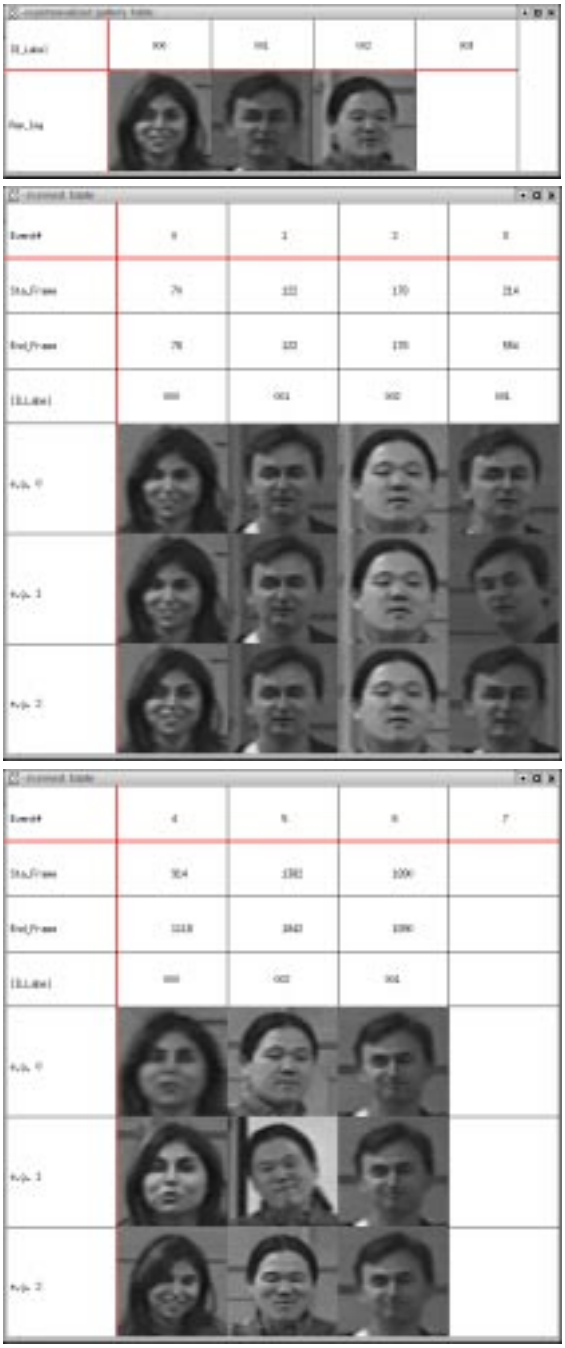

Figure 4: An example of the system's output. Top table displays the images that are most similar to the prototypes in the known-person database. Middle and bottom tables display personal events extracted from the input stream in each column. The first row of the tables describes the event number; 2nd: the starting frame number; 3rd: the ending frame number; 4th: the identity label; 5th: the facial image of the starting frame; 6th: the most similar face to the prototype; 7th: the facial image of the ending frame.

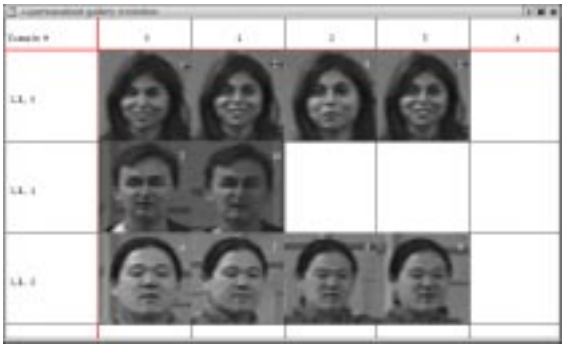

Figure 5: Temporal variation of the known-person entries. The variation is visualized by showing an image most similar to an entry's prototype at a time when the prototype was revised.

Figure 5 illustrates the temporal variation of each entry in the known-person database. Each image in the figure represents the prototype of an entry at the time the prototype was updated by showing the image most similar to the new prototype. The numbers within the images indicate the order of the updates over time. The figure shows our system's capability to adapt its known-person knowledge to temporally varying personal appearance under image variations, such as facial expressions.

We have also tested this system with several other test sequences including more illumination variations (turning a room-light on and off) and audiences (moving between speaker and camera). Speakers were correctly recognized except in a few cases in which two personal database entries were generated for a single person due to the illumination variations.

4 Discussion

We have described a framework for integrating the processes of object recognition and knowledge adaptation, as well as its application to the task of automatic video indexing based on personal appearance events. Our experimental results show successful performance of our system with a short video stream recorded off-line and with a variety of image variations.

The integration of the two processes provides a number of practical advantages. For example, such integration provides a basis for automating the data collection process required for many learning systems, which is often done manually and usually labor-intensive. Moreover, our framework can serve as a sensible solution to the problem of recognizing a person we have not seen for many years. This task is difficult even for humans since the effects of aging may change personal appearances considerably. State-of-

the-art face recognition systems still cannot solve this problem completely, given the inherent difficulty of the task [2, 5]. Our approach appear to improve identification performance, by keeping up to date a known-person database, instead of matching a current picture to ones captured many years ago. Furthermore, the framework is qualitatively more biologically plausible than classical pattern recognition systems. This may lead us to many other attractive application scenarios, such as a lifetime learning system.

As future work, our system's performance needs to be validated with longer video streams with more people and image variations. In order to realize this, we plan to merge the above-described system with our real-time face recognition system [9] in the near future. Currently, the decision function of our system is based on static and sequentially applied threshold functions. Learning these functions from experience is required for making our system more autonomous.

The use of the torso-color cue in our system is intuitive but is not, in fact, an optimal choice; it assumes that one person dresses the same way over time. Over a long range of time, this assumption becomes obviously invalid. However, the use of only the facial cue does not provide sufficient information when, for example, the size of a facial image is very small. We will consider adding additional visual or auditory cues to our system for improving performance and making the system more flexible. This extension, however, poses another problem: the *sensor fusion problem*, or how to merge the analysis results of different cues into a single decision. Our future work will address this problem.

Note that our system is constrained in that it may not recognize faces with head poses other than frontal. The capability to account for different head poses assumes greater importance in practical applications. Our previous work of modeling pose-varying facial images by parametric piecewise linear subspaces [4] can be incorporated into the proposed framework, in order to make the system robust against head pose variations.

For an automatic video indexing system, there are many types of event other than the personal appearance event described herein that can be processed. Additional modules, which analyze various events, should be added to our system for realizing a more flexible and complete indexing system.

Acknowledgments

This work was partially supported by a research grant from FXPAL, Palo Alto, USA. The authors wish to thank developers of the FLAVOR libraries for providing this work's software platform.

References

[1] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83:705–740, 1995.

[2] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg. The bochum/usc face recognition system: And how it fared in the feret phase iii test. In *Face Recognition: From Theory to Applications*, pages 186–205. Springer-Verlag, 1998.

[3] K. Okada and C. von der Malsburg. Automatic video indexing with incremental gallery creation: Integration of recognition and knowledge acquisition. In *Proceedings of the International Conference on Knowledge-Based Intelligent Information Engineering Systems*, pages 431–434, 1999.

[4] K. Okada and C. von der Malsburg. Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method. submitted to the IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[5] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[6] J. Piaget. *The Construction of Reality in the Child*. Ballantine, 1954.

[7] S. Satoh and T. Kanade. Name-it: Association of face and name in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 368–373, 1997.

[8] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 775–781, 1997.

[9] J. Steffens, E. Elagin, and H. Neven. Personspotter - fast and robust system for human detection, tracking and recognition. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 516–521, 1998.

[10] F. Tomita and Tsuji S. *Computer Analysis of Visual Textures*. Kluwer Academic Publishers, 1990.

[11] J. J. Weng and W. S. Hwang. Towards automation of learning: The state self-organization problem for a face recognizer. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 384–389, 1998.

[12] L. Wiskott, J.-M. Fellous, N. Krueger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.