Automatic Video Indexing with Incremental Gallery Creation: Integration of Recognition and Knowledge Acquisition

Kazunori Okada[†,‡] **and Christoph von der Malsburg**[†,§]
[†] Computer Science Department, University of Southern California, USA
[‡] Human Information Research laboratory, ATR, Japan
[§] Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
{kazunori,malsburg}@selforg.usc.edu

Object recognition plays a crucial role in our visual system by associating sensory inputs to the internal knowledge about known objects stored in memory. This association provides us with information required to interact with the environment. Another important aspect of the visual system is learning: autonomous knowledge acquisition from raw sensory data. Newly encountered objects need to be added to the previously acquired internal knowledge. Furthermore, since appearances of objects may change continuously, the internal knowledge about the objects has to be incrementally updated in order to maintain accurate representations.

Our investigation focuses on the fact that processes of object recognition and knowledge acquisition are not independent of each other. The state of adaptive 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. This interdependency suggests that these two processes need to be modeled together in a single framework. In computer vision, the task of object recognition and knowledge acquisition has often been treated independently. Most previous systems for example-based object recognition have treated the internal knowledge of objects as a *static* object gallery which was generated manually [1]. Thus the performance of these systems relies on the specific gallery that the developers chose to use.

The long-term goal of our research is to *integrate* object recognition and knowledge acquisition into a single example-based architecture introducing a *dynamic* relation between the performance and the state of internal knowledge. In this architecture, a system learns while performing; the internal knowledge about objects can be incrementally adapted from experiences, in on-line fashion, so that the performance of the recognition process will remain robust against temporal variations of object appearances. A direct and practical advantage of this integration is the automation of the gallery creation, which was usually done manually in previous studies. Weng and Hwang [5] recently proposed an on-line incremental learning system for the task of face recognition. Their system is based on statistical approximation methods such as PCA and LDA for recursive partitioning of the input feature space. These statistical approaches are usually time-consuming and requires recomputation of the internal model each time a new sample is added. Example-based approach simplifies the implementation of incremental learning since the previously acquired knowledge can be modified simply by addition or subtraction of samples.

In order to illustrate the proposed architecture, we developed a prototype of an automatic video indexing system. The task of video indexing takes a video stream as input and extracts events from it by spatiotemporal segmentation. These extracted events can serve as symbolic indices of a visual database, which can be used to reduce the search-time complexity of the database. In general, the definition of these events includes a wide variety of objects and their behavioral states [3]. In this study, we concentrate on an event of *personal appearance* which provides information of *who* appears *when* in an input scene. Satoh and Kanade [2]

demonstrated a technique for indexing facial identities by associating cooccurrence 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 and assumed a static gallery of names.

In the first stage of our system, we use facial information to segment an input sequence in space and time; a combination of multiple cues is used to extract spatiotemporal clusters of faces from the input. For spatial segmentation, facial regions within each frame are detected by a coarse-to-fine search using motion, convex shape and facial similarity cues [4]. A time-discontinuity cue of the facial movement trajectory is used for temporal segmentation. This spatiotemporal segmentation results in a set of sub-sequences, each of which contains only the face of *one* person.

In the second stage, two processes take place simultaneously: 1) the estimation of an identity from an input sub-sequence and 2) the adaptation of a personal gallery according to the results of this estimation process. Each input sub-sequence in this stage is represented by two cues: a sequence of a) Gabor-jet based facial representations [6] and b) color histograms of the torso in each frame of the input. The identity of the input is first estimated by a nearest neighbor search with the facial similarity cue. A confidence value resulting from this process determines whether the input belongs to a known person or if it requires further analysis. In the latter case, the torso-color cue is used to determine whether the input belongs to a known or unknown person. Thus the torso-color cue is used only when the facial similarity cue cannot provide sufficient information. The personal gallery is adapted by 1) updating an entry of the gallery with the input (known person case) or 2) adding the input as a new entry of the gallery (unknown person case). We present some experimental results of applying our system to scenes from a podium speech setting with freely moving speakers. These results demonstrate the system's potential for correct indexing of personal identities and for the automatic generation of a personal gallery.

References

[1] Rama Chellappa, Charles L. Wilson, and Saad Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[2] Shin'ichi Satoh and Takeo Kanade. Name-it: Association of face and name in video. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 368–373, 1997.

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

[4] Johannes Steffens, Egor Elagin, and Hartmut Neven. Personspotter - fast and robust system for human detection, tracking and recognition. In *Proceedings of the International Conference on Face and Gesture Recognition*, pages 516–521, 1998.

[5] 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 Face and Gesture Recognition*, pages 384–389, 1998.

[6] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krueger, and Christoph von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 19:775–779, 1997.