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
[§] Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
{kazunori,malsburg}@selforg.usc.edu

keywords: video indexing, adaptive known person database,
face recognition, spatiotemporal segmentation

Abstract

A framework for integrating the processes of object recognition and knowledge acquisition is proposed and applied to solve a task of automatic video indexing based on personal appearance events in a video stream. Spatiotemporal segmentation using multiple cues and example-based adaptation of a known person gallery are combined in a prototype system which demonstrated successful results in our preliminary experiments.

1 Introduction

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 [8] 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 require recomputation of the internal model 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 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 [5]. 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 [4] 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 ad-

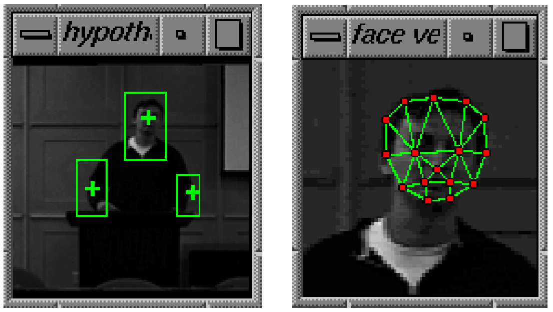


Figure 1: Examples of the spatiotemporal segmentation process. Left: ROIs found by the motion and the convex shape cues, Right: result of a bunch graph matching with a coarse graph.

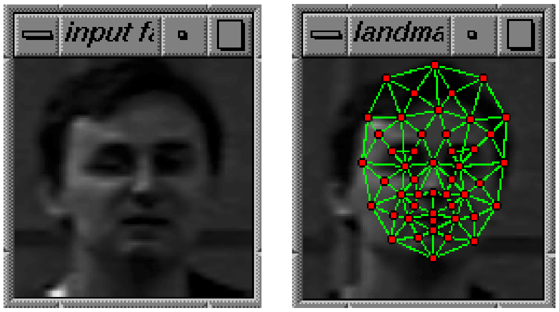


Figure 2: Examples of generating a facial representation. Left: cropped and normalized input frame, Right: result of a bunch graph matching with a fine graph.

dress, however, the issue of incremental knowledge acquisition based on visual information and assumed a static gallery of names. Our system extracts events of personal appearances by *spatiotemporal* segmentation of an input sequence. Each segmented person is then recognized from a *personal gallery* which is continuously adapted by the results of this recognition process. The rest of this paper describes this system and results of preliminary experiments.

2 The System

2.1 Spatiotemporal Segmentation

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 [6]. The motion and convex shape cues are first applied to each frame, resulting in a set of region of interests (ROIs). Each ROI is then cropped and normalized to a fixed size (128×128 pixels). We perform a bunch graph matching [9] on these normalized ROIs with a coarse graph (16 nodes). A similarity value between the ROI and the bunch graph is used as a confidence measure for the presence of a frontal face. A threshold function is applied to this confidence value in order to determine whether a given ROI contains a face. Figure 1 illustrates an example of the result of this process.

A time-discontinuity cue of the facial movement trajectory is used for temporal segmentation. The facial movement trajectory connects smoothly moving faces in consecutive frames. When a new face is found in a frame, it is kept track of in the following frames. The trajectory is discontinued when 1) no face is found within a current frame or 2) a spatial displacement of a face between two consecutive frames exceeds a proximity threshold (thus violating a smoothness constraint). This spatiotemporal segmentation results in a set of sub-sequences, each of which contains only the face of *one* person. These sub-sequences serve as inputs for the next stage.

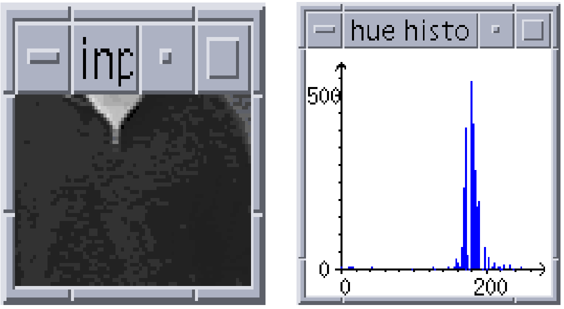


Figure 3: Examples of generating a torso-color representation. Left: cropped region of the torso, Right: histogram of pixel values (only the hue histogram is shown).

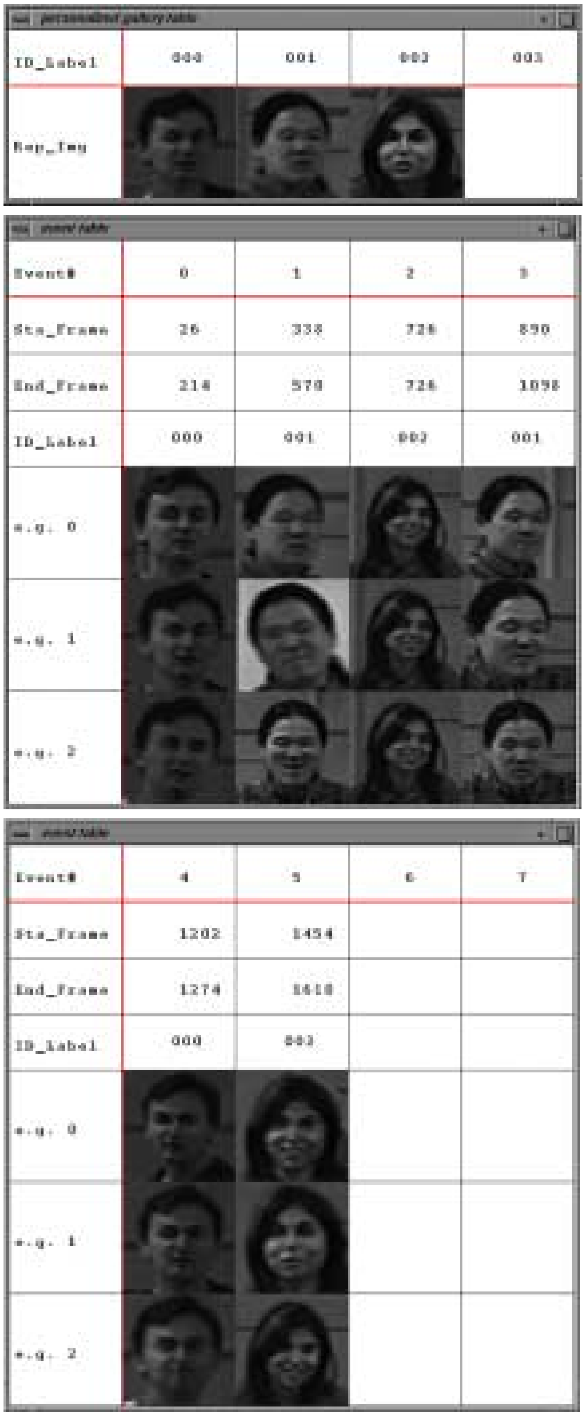
2.2 Identification and Gallery Adaptation

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 [9] and of b) color histograms of the torso in each frame of the input. Each frame of the input sub-sequence is subjected to a bunch graph matching with a fine graph (48 nodes), resulting in a sequence of the facial representations (see figure 2). A region of the torso is cropped from an original color frame at a location determined heuristically from the corresponding facial position. For the torso region in each frame, histograms of color pixel values are computed separately for each field of the Hue-Saturation-Intensity (HSI) color space (see figure 3). This representation of a single frame by the pair of a facial representation and a set of torso-color histograms (one for each color field) is called a *view representation*. A personal gallery contains a set of entries, each of which represents a single person. Each entry consists of a prototype and a set of view representations. The prototype is defined as an average of all the view representations in the entry. Thus it consists of an averaged facial representation and a set of averaged torso-color histograms.

The identity of an input sub-sequence is first estimated by a nearest neighbor search with the facial similarity cue. Only when the facial similarity cue cannot provide sufficient information, is the torso-color cue used. A similarity value between two facial representations is computed by averaging the normalized dot-products of two corresponding local feature vectors (jets) on every node of the facial graph. The input is compared to a gallery entry by averaging the facial similarities between each facial representation of the input and a prototype of the entry. The entry with the highest average similarity is chosen as a candidate and its similarity is used as a confidence measure which is subjected to a threshold function. If the confidence value exceeds a threshold, the input is identified as this candidate. Otherwise, the decision is deferred to a next step with the torso-color cue. A similarity value between two torso-color representations (HSI histograms) is given by an average of Kolmogorv distances of the Hue and Saturation histograms [7]. The Intensity histogram is not used in order to mitigate influences from illumination variations. Torso-color similarity between the input sub-sequence and the candidate's prototype is computed by averaging the similarities between each torso-color representation of the input and a prototype of the candidate. This average torso-color similarity is subjected to another threshold function. When the torso-color of the input is very similar to the one of the candidate, the input is identified as the candidate. Otherwise, it is recognized as a previously unknown person. The threshold for the torso-color similarity is set with a qualitatively more strict criterion than for the facial similarity, in order to reduce a risk of false identifications caused only by similar torso-color.

The personal gallery is modified each time an input is recognized. When an input is identified as a known person, a sequence of view representations from this input is added to a corresponding gallery entry and the prototype of this entry is updated by recomputing new average representations. When the input is recognized as an unknown person, a new entry is generated from this input and added to the personal gallery as a new person.



ID_Label	000	001	002	003
Rep_Img				

Event#	0	1	2	3
Sta_Frame	26	338	726	890
End_Frame	214	570	726	1098
ID_Label	000	001	002	001
e.g. 0				
e.g. 1				
e.g. 2				

Event#	4	5	6	7
Sta_Frame	1202	1454		
End_Frame	1274	1610		
ID_Label	000	002		
e.g. 0				
e.g. 1				
e.g. 2				

Figure 4: Examples of the system's output. Top table displays a personal gallery automatically generated from a test input sequence. Middle and bottom tables display personal events extracted from the test input in each column. First row of the tables describes event number, 2nd: starting frame number, 3rd: ending frame number, 4th: identity label of this event, 5th: facial view of the starting frame, 6th: most similar face to the prototype, 7th: facial view of the ending frame.

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 would create 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 the system's output with a test sequence. This test sequence was captured by a cam-corder in a seminar room environment. It consists of 1800 frames and contains three speakers, speaking at a podium, moving freely, and reappearing to a podium occasionally. Our system successfully extracted six personal appearance events from this input, identified each person correctly, and automatically generated a correct personal gallery. Note that frame gaps between the personal appearance events corresponds to situations, where no speaker was present in a frame of view or the rotation of a speaker's head was very large. An average throughput of this system was one frame per second on a SGI workstation with a R10000 processor. We have tested this system with several other test sequences including more illumination variations (turning a room-light on and off) and audiences (moving audiences between a speaker and a camera). Speakers were correctly recognized except a few cases, in which two personal gallery 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 acquisition, as well as its application for the task of automatic video indexing based on personal appearance events. The integration of the two processes provides not only a basis for biological plausibility but also some practical advantages. For example, recognizing a person we have not seen for many years is a difficult task even for humans since the effects of aging may change personal appearances considerably. The state-of-art face recognition systems still cannot solve this problem completely [2, 3]. Our approach may serve as a sensible solution to this problem by automatically maintaining a personal gallery up-to-date, instead of matching a current picture to ones captured many years ago, which is inherently a difficult task.

The Results of the preliminary experiments presented in this paper is encouraging although our system is based on very simple implementation designs. Additional tests with more difficult variations and longer input sequences are needed for further development of the system.

The use of the torso-color cue in our system is intuitive but would not be an optimal choice; it assumes that one person dresses the same way over time. Over a long range of time, this assumption may become impractical. However, to use only the facial cue would not provide sufficient information when the size of a facial image is very small (see figure 1) or front views of a face are not available. Therefore, additional visual or auditory cues may be incorporated to improve our system. The decision function used in the second stage of our system is based on a sequence of simple threshold functions which require a parameter tuning. Learning these functions from experiences would be another future topic. Furthermore, as a video indexing application, our system could be extended by adding modules that analyze events other than the personal appearance.

Acknowledgments

This work was supported by a research grant from FXPAL, Palo Alto, USA. The authors wish to thank Thomas Maurer for his helpful advice and Hai Hong for his contribution towards the development of this system.

References

[1] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):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] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation. In *Face Recognition: From Theory to Applications*, pages 244–261. Springer-Verlag, 1998.

[4] S. Satoh and T. 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.

[5] M. A. Smith and T. 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.

[6] 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 Face and Gesture Recognition*, pages 516–521, 1998.

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

[8] 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.

[9] 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.