PART III

Chapter 13

Analyzing Dynamic Faces: Key Computational Challenges

By Pawan Sinha

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Correspondence to be addressed to: psinha@MIT.EDU

Part III – Chapter 13
Analyzing Dynamic Faces: Key Computational Challenges
By Pawan Sinha

Introduction

Research on face perception has focused largely on static imagery. Featural details and their mutual configuration are believed to be the primary attributes subserving tasks such as identity, age and expression judgments. The extraction of these attributes is best accomplished in high-quality static images. In this setting, motion is counter-productive in that it complicates the extraction and analysis of details and spatial configuration. However, a counterpoint to this idea has recently begun emerging. Results from human psychophysics have demonstrated that motion information can contribute to face perception, especially in situations where static information, on its own, is inadequate or ambiguous. This body of work has served as an impetus for a computational investigation of dynamic face analysis. The chapters included in this volume are excellent examples of the kinds of issues and approaches researchers are exploring in this domain. Along these lines, my intent in this article is to highlight some of the key computational challenges that an analysis of dynamic faces entails. This is by no means an exhaustive list, but rather a set of issues that researchers are likely going to have to tackle in the near term.

When (and what) does motion contribute to face perception?

Although we said above that motion contributes to face perception, it has to be acknowledged that we do not yet have a good characterization of the tasks which benefit significantly from the inclusion of dynamic information. As we all have experienced first hand, several face perception tasks can be accomplished quite well even with purely static images. What exactly is motion good for? Does it make a qualitative difference in the performance of certain face tasks or is it only a 'bit player'? Human experimental literature has not so far provided a clear answer to this question. One has to work hard to design stimuli where the contribution of motion is clearly evident. Given the equivocal picture from the experimental front, it falls upon computational investigations to help identify task domains which are likely to benefit in significant ways by the availability of dynamic information. For instance, if it can be shown that under some simple choices of features and classifiers, dynamic attributes of expressions are much more separable than their static manifestations, then one may justifiably predict that human performance on the task of expression classification will be significantly facilitated by motion information. More generally, the idea would be to build a taxonomy of face perception tasks based on a computational analysis of how much task-related information is carried by static and dynamic facial signals. This endeavor would lead to an exciting interplay between computational researchers and psychophysicists, with the former actively suggesting promising experimental avenues to the latter.

Besides characterizing the tasks that motion might contribute to, it is also important to investigate precisely what kind of information motion is adding to the computation. Are facial dynamics useful primarily for estimating three-dimensional structure, or for performing a kind of super-resolution to effectively gain more detail information beyond that available in any single frame, or simply for figure-ground segregation. These are, of course, not the only possibilities. But the larger point is that we need to understand how motion might come into play during a face-perception task. Computational simulations can play a valuable role in this investigation by providing indications of how feasible it is to extract different attributes (3D shape, high-resolution images, figure-ground relations etc.) from real-world video data.

How can we capture facial dynamics?

The front end for a static face analyzer is fairly straightforward: a camera that can take a short exposure snap-shot and a program that can detect fiducial points such as the centers of the pupils and corners of the mouth. With dynamic imagery, the analogous task becomes much more challenging. Human assistance in annotation, which is a feasible option with static images, is no longer realistic with dynamic sequences on any significant length. Researchers and practitioners in the applied domain of facial motion capture have turned to the use of crutches like grids of painted dots or reflective stickers. While this simplifies the problem to an extent, it is not a full solution for at least two reasons. First, it is not always possible to attach such markers to faces (imagine trying to perform motion capture with archival footage). Second, even when feasible, this approach provides only a sparse sampling of motion information across the face. Much of the nuance of facial movements is lost. This is evident in the unnatural dynamics of animated faces in the current crop of movies. What we need are computational techniques for obtaining dense motion information from unmarked faces.

Walder et al. in this volume describe an important step forward in this direction. Their algorithm takes as its input an unorganized collection of 4D points (x, y, z and t), and a mesh template. Its output is an implicit surface model which incorporates dense motion information and closely tracks the deformations of the original face. The results they present are striking in their fidelity. This work sets the stage for addressing the next set of challenges in dynamic face tracking. An obvious one is the need to be able to work with 2D rather than 3D spatial information. For human observers, a 2D video sequence typically suffices to convey rich information about facial dynamics. A computational technique ought to be able to do the same. This is important not merely to mimic the human ability, but also from the practical standpoint of being able to work with conventional video capture systems, the vast majority of which are 2D. Perhaps a combination of past 3D estimation techniques developed in the context of static face analysis (Blanz and Vetter, 1999), and the kind of approach described by Walder et al. can accomplish the goal of motion capture from 2D sequences. Blanz and Vetter's technique allows for the generation of 3D models from single 2D images based on previously seen 2D-3D mappings. Once such a 3D structure is estimated, it can be used to initialize the stimulus-to-template alignment in Walder's approach. It remains to be seen whether an initial 3D estimation step will suffice for motion tracking over extended sequences, or if the 3D estimation will need to be repeated at frequent intervals for intermediate frames.

A second important avenue along which to push the motion tracking effort is working with poor-quality video. Besides yielding obvious pragmatic benefits, an investigation of how to handle low spatial and temporal resolution video is likely to be useful in modeling human usage of dynamic information. Although static facial information appears to be sufficient for many tasks when the images are of high-resolution, the significance of dynamic information becomes apparent when spatial information is degraded. Johansson's classic displays, and their more recent derivatives, are a testament to this point. Computationally, however, the derivation of dense motion fields from low-resolution videos presents significant challenges. It is hard to establish spatially precise correspondences in such sequences, and hence the accuracy and density of the recovered motion fields is limited. Human observers, however, are quite adept at this task. What kind of a computational strategy can prove robust to spatio-temporal information degradation? One possibility is the use of high-resolution internal models that can be globally fit to the degraded inputs in order to establish more precise local correspondences. This general idea of using internal models that are richer than the inputs is very similar in spirit to what we described above for working with 2D rather than 3D data. Perhaps this approach will prove to be a broadly applicable strategy for handling many different kinds of information loss in the observed facial sequences.

How can dynamic facial information be represented?

Having tracked a dynamic face using the kinds of approaches outlined above, the next computational challenge is to efficiently represent this information. The different appearances of a face that are revealed over the course of tracking together constitute its temporally extended appearance model, or TEAM for short, as illustrated in figure 13-1.

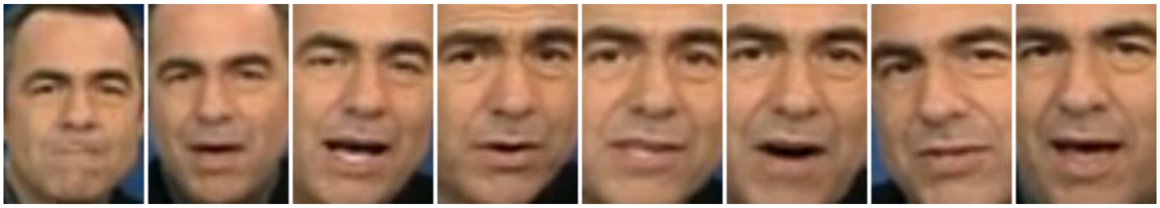

Figure 13-1. *The result of facial tracking is a highly redundant 'stack' of images that we refer to as a TEAM, for "Temporally Extended Appearance Model." How TEAMs should be encoded for various face-perception tasks is an important open problem.*

The resulting spatio-temporal volume of face appearances constitutes a rich data set for constructing a robust face model, but how exactly should we encode it? This apparently simple question has yet to be satisfactorily answered either experimentally or computationally. Dynamic face representation is the fundamental challenge that the chapters by Boker & Cohn and Serre & Giese in this volume expound on. Here, we shall outline the key conceptual issues related to this topic.

The simplest thing we could do with a spatio-temporal volume is store it in its entirety for future reference. Individual images or new spatio-temporal inputs could be compared to the stored volume using any set of features we wish. The storage requirements of this strategy are obviously prohibitive, though. If we are to remember every spatio-temporal face volume we encounter, we will be overwhelmed with data very quickly. Even discarding the temporal contingencies between frames and maintaining only newly encountered static images does not do much to mitigate this very expensive encoding strategy. To learn anything useful from dynamic data, the visual system must represent spatio-temporal volumes efficiently and with enough information for recognition to proceed.

There is a great deal of redundancy in the data depicted in Figure 1, and finding encodings that reduce it can help guide the search for an efficient representation. However, there are two issues we must be mindful of as we consider possible methods of redundancy reduction in this setting: First, does a particular representation support robust recognition? Second, is the proposed representation consistent with human psychophysical performance?

To consider the first issue, there are existing methods for recovering a “sparse” encoding of natural image data (van Hateran and Ruderman, 1998; Olshausen, 1996, 2003). Implementing these methods on local image patches or spatio-temporal sub-volumes tends to produce basis functions resembling edge-finding filters which translate over time. These provide a useful vocabulary for describing an image or an image sequence using a small set of active units, but these features are often not ideal for recognition tasks (Balas and Sinha, 2006).

In terms of our second issue, building representations that are consistent with human performance, there are many computations we could carry out on our volume which are “too strong” in some sense to be relatable to human performance. For example, we could potentially use our image sequence to reconstruct a 3-D volumetric model of the observed face using structure-from-motion algorithms (Ullman, 1979). The smoothness of appearance change across the images in a TEAM could reduce the usual difficulty of solving the correspondence problem, and we can easily obtain far more object views than strictly necessary to solve for 3D form. However, faces do not respect a cornerstone of structure-from-motion computations, namely, rigidity. The non-rigid deformations that a face typically goes through make it difficult to estimate its 3D structure from a TEAM stack. Furthermore, it seems unlikely that human observers actually recognize faces based on view-invariant volumetric models of shape (Ullman, 1996). View-based models currently seem more commensurate with the psychophysical data (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992). However, to revisit a point raised earlier, there are also good psychophysical reasons against storing all the views of an object within some spatio-temporal volume. Specifically, observers trained to recognize novel dynamic objects do not behave as though they have stored all the views they are trained with. For example, they find novel dynamic objects harder to recognize if the stimulus presented at test is the time-reversed version of the training sequence (Stone, 1998, 1999; Vuong and Tarr, 2004). An ideal observer that maintains a representation of each image should not be impaired by this manipulation, suggesting human observers do not simply store copies of all object views encountered during training. Instead, the order of appearances is maintained and becomes a critical part of the representation.

Learning purely local features in space and time is useful within particular domains (Ullman and Bart, 2004) but potentially difficult to “scale up” to natural settings. Also, maintaining fully volumetric face models or large libraries of static face views is both inefficient and inconsistent with human data. The challenge we are faced with then is to develop a compact and expressive representation of a dynamic face that is consistent with human performance.

A face-model based on temporal association might offer an attractive alternative to existing proposals. Instead of storing an intractably large number of ordered static views, it should be possible to store only a few prototypical images and use dynamic input to learn a valid generalization function around each prototype. Redundancy reduction within a spatiotemporal volume is thus accomplished at the level of global appearance (however we choose to represent it) and the ultimate encoding of the face is view-based with a learned "tuning width" in appearance space around each prototype view.

There are multiple aspects of this model that have yet to be thoroughly explored psychophysically. For example, how are prototypical views of a face selected within a volume? There is very little work on how such views (or "keyframes" (Wallraven and Bülthoff (2001)) might be determined computationally or the extent to which they are psychologically real. Likewise, we do not yet have a detailed picture of how generalization around an image evolves following dynamic exposure. We have recently suggested that distributed representations of object appearance follow from dynamic experience with a novel object, but as yet we have not investigated the long-term consequences of dynamic training. These two issues constitute key parameters in what is essentially a statistical model of dynamic object appearance. Finding "keyframes" is analogous to identifying multiple modes in the data, while understanding patterns of generalization around those keyframes is analogous to identifying the variance of data around some mode. In this framework, motion is not a new feature for recognition, but rather a principled way to establish a sort of "mixture model" for face appearance.

The advantage of this strategy is that it makes explicit the fact that while observers probably have access to global appearance data, temporal data is only available locally. Thus, we do not try to build a face representation that covers the whole viewing sphere of possible appearances (Murase and Nayar, 1995). Instead, we limit ourselves to learning what changes a face is likely to undergo in a short time interval. This basic proposal leads to many interesting questions for psychophysical research, and makes easy contact with several physiological studies of object representation in high-level cortical areas.

To summarize, the question of dynamic face representation has not yet been adequately explored in the experimental or computational domain. In the absence of human data to guide computational strategies, current proposals are necessarily speculative. One idea that seems perceptually plausible and computationally attractive is to encode a TEAM via 'keyframes' and some specification if the transformation linking these keyframes. Keyframes can be computed via a cluster analysis. They would correspond to the frames that minimize the sum of distances between themselves and all other frames of the TEAM, under the constraint of minimizing the number of keyframes. Of course, the error metric will keep decreasing monotonically as more and more keyframes are selected. However, as is the case with principal components analysis, the decrease in error obtained by adding a new keyframe diminishes as the number of key frames increases. The knee of the corresponding screen plot would indicate the number of keyframes to be included. As for encoding the transformation linking these keyframes, a manifold in a low-dimensional space, for instance, one corresponding to the principal components of object appearances seen, might be adequate. The computational choices here await experimental validation.

What aspects of motion information are important for specific face-perception tasks?

If we can convince ourselves that motion information does indeed play a significant role in face perception, a more specific question becomes pertinent: Precisely which aspects of the overall motion signal contribute to a given face-perception task? As a rough analogy, consider the case of static facial analysis. We know that photometric information plays a role in several facial judgments such as those pertaining to identity and aesthetics. However, this is too broad a statement to be interesting or useful. It needs to be made more precise; which aspects of the photometric signal are really crucial for, say, identification? Computational and experimental results point to the ordinal brightness relationships around the eyes as being of key significance (Viola and Jones, 2001; Gilad et al., 2009). A similar kind of investigation is needed in the dynamic setting.

The computational challenge here is to consider many possible subsets of the full dynamic signal in order to determine which ones are the most useful for the performance of a given task. For the case of identification, for instance, is the movement of the mouth more discriminative across individuals than the movement of the eyes? The chapter by Bartlett et al. presents an excellent instance of this kind of an effort applied to the domain of expression perception. Their work demonstrates how such an approach can reveal hitherto unknown facial attributes as indicators of subtle differences in mental states such as engagement and drowsiness.

It would indeed be interesting to examine whether the kinds of feature saliency hierarchies that have been determined for static faces (Fraser et al., 1990) also carry-over to the dynamic setting, or whether the two sets are entirely distinct. Besides carving up information spatially, it will also be important to consider subsets of the dynamic signal in the spatio-temporal frequency domain. Just as static face analysis appears to depend most strongly on a constrained band of spatial frequencies (Costen et al., 1996), so might dynamic face analysis be driven largely by a small subset of the full spatio-temporal spectrum. It may be the case, for instance, that seemingly small flutters of eyelids might be more informative for some face-perception tasks than large-scale eye-blinks.

A related, but conceptually distinct issue is that of the duration of motion information necessary for performing different face tasks to some specified level of accuracy. Here, computational simulations can prove to be very useful in studying how a system's performance declines as the length of the motion sequence it is shown is made shorter and shorter. Not only would this provide benchmarks for psychophysical tests of human performance (and for ideal-observer analyses), it would also have the important side-effect of suggesting hypotheses for the first question we mentioned above: Which aspects of motion are important for a given task? Some motion signals might be ruled out as significant contributors just based on the fact that a viewing duration might be too short for that kind of motion to occur.

Conclusion

The analysis of dynamic faces is an exciting frontier in face-perception research. The terrain is wide open and several of the most basic questions remain to be answered, both from an experimental perspective and a computational one. We have attempted here to list a few of these questions. We hope that the coming months and years will see a shift in the field's focus from purely static imagery to more realistic dynamic sequences. The chapters in this volume represent exciting initial steps in this direction.

Bibliography

Balas, B. & Sinha, P. (2006). Receptive Field Structures for Recognition. Neural Computation 18, 497-520.

Blanz, V. and Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. SIGGRAPH Conference Proceedings, pp 187-194.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical Support for a 2-Dimensional View Interpolation Theory of Object Recognition. Proceedings of the National Academy of Sciences of the United States of America, 89(1), 60-64.

Costen, N. P., Parker, D. M. & Craw, I. (1996) Effects of high–pass and low-pass spatial filtering on face identification. Percept. Psychophys. 58, 602-612.

Fraser, I. H., Craig, G. L. & Parker, D. M. (1990) Reaction time measures of feature saliency in schematic faces. Perception 19, 661-673.

Gilad, S., Meng, M. and Sinha, P. (2009). Role of ordinal contrast relationships in face encoding. Proceedings of the National Academy of Sciences, 106(13), 5353-5358.

Murase, H. & Nayar, S. K. (1995). Visual Learning and Recognition of 3-D Objects from Appearance. International Journal of Computer Vision, 14, 5-24.

Olshausen, B. A. (2003). Principles of Image Representation in Visual Cortex. In The Visual Neurosciences, L.M. Chalupa, J.S. Werner, Eds. MIT Press, 2003: 1603-1615.

Stone, J. V. (1998). Object recognition using spatiotemporal signatures. Vision Research 38, 947-951.

Stone, J. V. (1999). Object recognition: view-specificity and motion-specificity. Vision Research 39, 4032-4044.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. Cognit Psychol, 21(2), 233-282.

Ullman, S. (1979). The interpretation of structure from motion. Proceedings of the Royal Society of London, Series B 203, 405-426.

Ullman, S. (1996). High-Level Vision. MIT Press, Cambridge, MA.

Ullman, S. & Bart, E. (2004). Recognition invariance obtained by extended and invariant features. Neural Networks, 17, 833-848.

van Hateran, J. H. & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proceedings of the Royal Society of London, Series B, 265, 2315-2320.

Viola, P., & Jones, M. (2001). Rapid Object Detection Using a Boosted Cascade of Simple Features. Paper presented at IEEE Computer Vision and Pattern Recognition.

Vuong, Q. C. & Tarr, M. J. (2004). Rotation direction affects object recognition. Vision Research 44, 1717-1730.

Wallraven, C. & Bülthoff, H. H. (2001). Automatic acquisition of exemplar-based representations for recognition from image sequences. In Proceedings of CVPR.