Wallraven, C., Schwaninger, A., Schumacher, S., & Bülthoff, H.H. (2002). *Lecture Notes in Computer Science, 2525,* 651-660.

View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho

Christian Wallraven[1),*], Adrian Schwaninger[1), 2),*], Sandra Schumacher,[2)] Heinrich H. Bülthoff[1)]

[1)] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2)] Department of Psychology, University of Zürich, Switzerland

Abstract. For humans, faces are highly overlearned stimuli, which are encountered in everyday life in all kinds of poses and views. Using psychophysics we investigated the effects of viewpoint on human face recognition. The experimental paradigm is modeled after the inter-extra-ortho experiment using unfamiliar objects by Bülthoff and Edelman [5]. Our results show a strong viewpoint effect for face recognition, which replicates the earlier findings and provides important insights into the biological plausibility of view-based recognition approaches (alignment of a 3D model, linear combination of 2D views and view-interpolation). We then compared human recognition performance to a novel computational view-based approach [29] and discuss improvements of view-based algorithms using local part-based information.

1. Introduction

According to Marr [16] human object recognition can be best understood by algorithms that hierarchically decompose objects into their parts and relations in order to access an object-centered 3D model. Based on the concept of nonaccidental properties [14], Biederman proposed in his recognition by components (RBC) theory [1], that the human visual system derives a line-drawing-like representation from the visual input, which is parsed into basic primitives (geons) that are orientation-invariant. Object recognition would be achieved by matching the geons and their spatial relations to a geon structural description in memory. This theory has been implemented in a connectionist network that is capable of reliably recognizing line drawings of objects made of two geons [11].

In object recognition, view-based models have often been cited as the opposite theoretical position to the approaches by Marr and Biederman[1]. Motivated by the still unsolved (and perhaps not solvable) problem of extracting a perfect line drawing from natural images different view-based approaches have been proposed. In this paper we consider three main approaches: Recognition by alignment to a 3D representation

[1] However, it is interesting that Biederman and Kalocsai [3] point out that face recognition – as opposed to object recognition – cannot be understood by RBC theory mainly because recognizing faces entails processing holistic surface based information.

[15], recognition by the linear combination of 2D views [25], and recognition by view interpolation (e.g., using RBF networks [19]). What these approaches have in common is that they match viewpoint *dependent* information as opposed to viewpoint *invariant* geons.

The biological plausibility of these models has been investigated by comparing them to human performance for recognizing paper clip and amoeboid like objects [5,7]. In contrast to those stimuli, faces are highly overlearned and seen in a vast variety of different views and poses. Therefore, we were interested whether a) human face recognition shows similar effects of viewpoint and b) by which of these view-based approaches face recognition can be best understood. We then compared human recognition performance to another view-based framework, namely the feature matching approach based on the framework introduced in [29]. Based on the results we discuss the role of parts and their interrelationship from a view-based perspective in contrast to the models proposed in [1,11].

2. Psychophysical experiment on view-based recognition of faces

2.1 Participants, Method and Procedure

Ten right-handed undergraduates (five females, five males) from the University of Zürich volunteered in this study. The face stimuli were presented on a 17" CRT screen. The viewing distance of 1 m was maintained by a head rest so that the faces covered approximately 6° of the visual angle. Twenty male faces from the MPI face database [4] served as stimuli.

The experiment consisted of a learning and a testing phase. Ten faces were randomly selected as distractors and the other 10 faces were selected as targets. During learning, the target faces were shown oscillating horizontally ±5° around the 0° and the 60° extra view (see Figure 1). The views of the motion sequence were separated by 1° and were shown 67 ms per frame. The oscillations around 0° started and ended always with the +5° view, the oscillations around 60° started and ended always with the +55° view. Both motion sequences lasted 6 sec, i.e. 4 full back-and forth cycles. For half the faces the 0° sequence was shown first, for the other half of the faces the 60° sequence was shown first. The order of the ten faces was counterbalanced across the ten participants. After a short break of 15 min the learning block was repeated and for each face the order of the two motion sequences was reversed.

In the testing phase, the subjects were presented with static views of the 10 target and the 10 distractor faces. The faces were shown in blocks of 20 trials in which each face was presented once in a random order. The test phase contained 300 trials and each face was presented once in each of the 15 angles depicted in Figure 1. Each trial started with a 1000 ms fixation cross followed by the presentation of a face. Participants were instructed to respond as fast and accurately as possible whether the presented face had been shown in the learning face (i.e. it was a target) or whether it was a distractor by pressing the left or right mouse button. On each trial, the faces

were presented until the button press occurred. The assignment of buttons to responses was counterbalanced across participants.

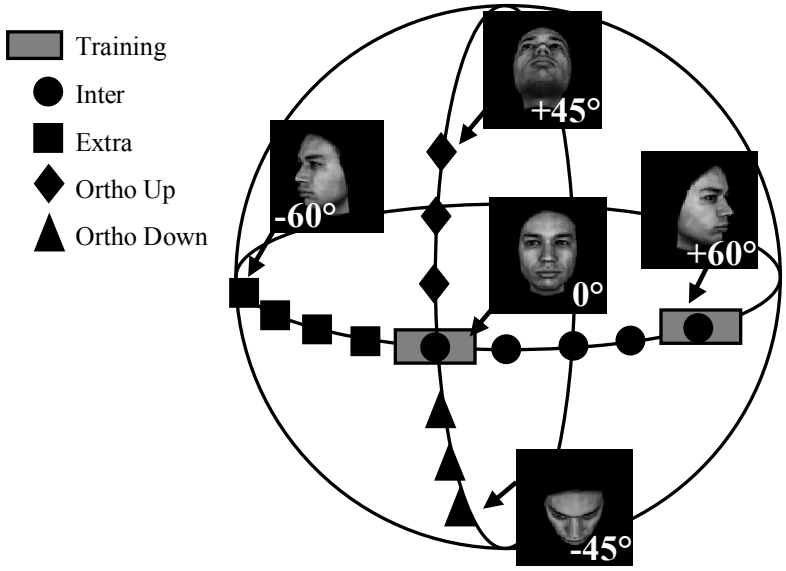

Figure 1. Training occurred at 0° ±5° (frontal view) and 60° ±5° (side view). Testing was performed for 15 views separated by 15°. The four testing conditions are labeled (inter, extra, ortho up, ortho down).

2.2 Results and Discussion

Signal detection theory was used to measure recognition performance. The relevant measure is d' = z(H) – z(FA), whereas H equals the hit rate, i.e. the proportion of correctly identified targets, and FA the false alarm rate, i.e. the proportion of incorrectly reporting that a face had been learned in the learning phase. H and FA are converted into z-scores, i.e. to standard deviation units. Individually calculated d' values were subjected to a two-factor analysis of variance (ANOVA) with condition (extra, inter, orthoUp, orthoDown) and amount of rotation (0, 15, 30, 45) as within subjects factors. Mean values are shown in Figure 2.

Recognition d' was dependent on the condition as indicated by the main effect of this factor, $F(3, 27) = 23.1$, $MSE = .354$, $p < .001$. There was also a main effect of amount of rotation, $F(3, 27) = 10.93$, $MSE = 1.500$, $p < .001$. The effect of rotation was different across conditions as indicated by the interaction between amount of rotation and condition, $F(9,81) = 3.30$, $MSE = .462$, $p < .01$. The four conditions were compared to each other using Bonferroni corrected pairwise comparisons. Recognition in the inter condition was better than in the extra condition ($p < .05$). Recognition in inter and extra conditions was better than in both ortho conditions ($p <$

.01). Finally, recognition performance did not differ in the two ortho conditions (p = .41). These results are difficult to explain by approaches using alignment of a 3D representation [15] because such a differential effect of rotation direction would not be expected. Moreover, human performance questions the biological plausibility of the linear combination approach for face recognition [25], because it cannot explain why performance in the inter condition was better than in the extra condition. The results can for example be understood by a linear interpolation within an RBF network [19] – in the next section, we present another view-based framework, which can model the results [29]. Both of these models predict inter > extra > ortho, which was shown clearly in the psychophysical data. Interestingly, the results of the present study lead to the same conclusions as the study in [5], who used paper clips and amoeboid objects in order to investigate how humans encode, represent and recognize *unfamiliar* objects. In contrast, in our study perhaps the most familiar object class was used. Thus, familiarity with the object class does not necessarily predict qualitatively different viewpoint dependence.

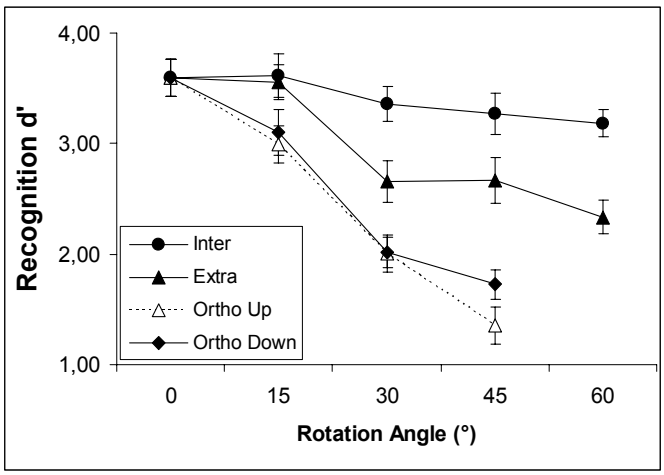

Figure 2. Human recognition performance in the four rotation conditions (inter, extra, ortho up, ortho down) across viewpoint (0° is the frontal view).

3. Computational modeling

3.1 Description of the system

The original inter-extra-ortho experiment was analyzed using radial basis function (RBF) networks, which were able to capture the performance of subjects in the various tasks (see also [18] for a study on face recognition using RBF networks). In this paper, we apply another kind of view-based computational model to the psychophysical data, which is based on a framework proposed in [29]. The motivation for the proposed framework came from several lines of research in psychophysics:

First of all, evidence for a view-based object representation - as already stated above - has been found in numerous studies (also from physiological research). In addition, recent results from psychophysical studies showed that the temporal properties of the visual input play an important role for both learning and representing objects [28]. Finally, results from psychophysics (see e.g., [13,21,22]) support the view that human face recognition relies on encoding and storing local information contained in facial parts (featural information) as well as the spatial relations of these features (configural information).

A model, which can incorporate elements of these findings, was proposed in [29]. The framework is able to learn extensible view-based object representations from dynamic visual input on-line. In the following, we shortly describe the basic elements of the framework as used in this study. Each image is processed on multiple scales to *automatically* find interest points (in our case, corners). A set of visual features is constructed by taking the positions of the corners together with their surrounding pixel patches (see Figure 4a). In order to match two sets of visual features, we consider an algorithm based on [20]. It constructs a pair-wise similarity matrix **A** where each entry A_{ij} consists of two terms:

$$A_{ij} = \exp(-\frac{1}{\sigma_{dist}^2}\mathrm{dist}^2(i,j)) \cdot \exp(-\frac{1}{\sigma_{sim}^2}\mathrm{sim}^2(i,j)) \qquad (1)$$

where $\mathrm{dist}^2(i,j)) = ((x_i - x_j)^2 + (y_i - y_j)^2)$ measures the distance between a feature pair and *sim(i,j)* measures the pixel similarity of the pixel patches (in our case, using Normalized Cross Correlation). The parameters $\sigma_{dist}, \sigma_{sim}$ can be used to weight distance and pixel similarity. Basen on the SVD of this matrix $\mathbf{A=UVW^T}$ we then construct a re-scaled matrix $\mathbf{A'=UW^T}$, which is then used to find a feature mapping between the two sets [20,29]. The goodness of the match is characterized by the *percentage of matches* between the two feature sets. This feature matching algorithm ensures that both global layout and local feature similarity are taken into account. It is important to note that there is neither a restriction to a global spatial measure in pixel space nor to a local measure of pixel similarity. Any kind of view-based feature measure can be introduced in a similar manner as an additional term in equation 1.

One of the advantages of this framework, which a purely view-based holistic representation lacks, is its *explicit* representation of local features. This enables the system amongst other things to be more robust under changes in illumination and occlusion [29,30]. Since the input consists of image sequences the visual features can also be augmented with *temporal information* such as trajectories of features. Temporal information is given in our case by the learning trials in which a small *horizontal* rotation is presented. We thus modified the distance term in equation 1 such that it penalizes deviations from the horizontal direction for feature matches by an increased weighting of the vertical distance between features *i* and *j*:

$$\mathrm{dist}^2(i,j)) = ((x_i - x_j)^2 + \alpha(y_i - y_j)^2) \quad \text{with } \alpha \geq 1 \qquad (2) \cdot$$

Figure 3 shows matching features[2] between two images for two settings of α=1 and α=3: α=1 (Figure 3a) yields a matching score of 30 percent, whereas α=3 (Figure 3b) yields a matching score of 37 percent. The rationale behind using the penalty term not

[2] Some matches between features are not exactly horizontal due to localization inaccuracies inherent in the corner extraction method.

only comes from the dynamic information present in the learning phase, but is also motivated by the psychophysical results in [5,7], where humans showed a general tendency towards views lying on the horizontal axis.

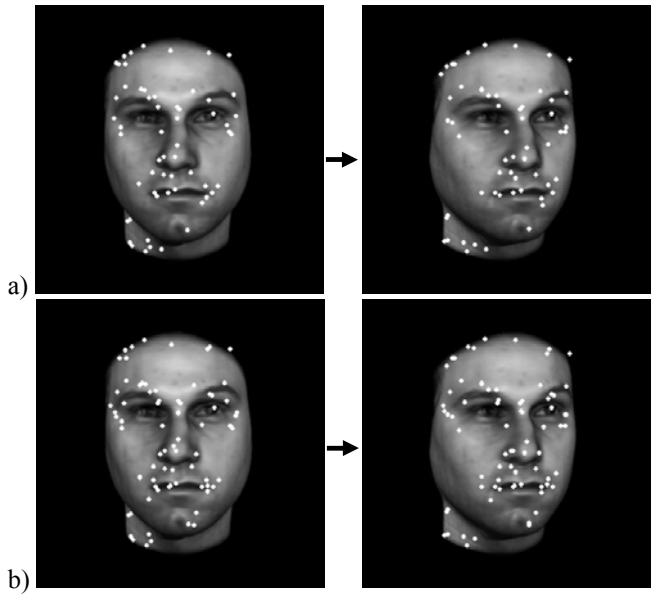

Figure 3. Matching features between two views of a face a) without vertical penalty term b) with vertical penalty term.

3.2 Computational recognition results

In the following, we present recognition results, which were obtained using the same stimuli as for the human subjects. Again, the system was trained with two small image sequences around 0° and 60° and tested on the same views as humans. The final learned representation of the system consisted of the 0° and 60° view, each containing around 200 local visual features. Each testing view was matched against all learned views using the matching algorithm outlined above. To find matches, a winner-takes-all strategy was employed using the *combined matching score* of the two learned views for each face. The results in Figure 4b show that our computational model exhibits the *same* qualitative behavior in performance as human subjects replicating the drop in performance, i.e. inter>extra>ortho[3]. Inter performance was best due to support from two learned views as opposed to support only from the frontal view for the extra conditions. Recognition of ortho views was worst due to three factors: first, inter conditions had support from two views, second, the learned

[3] The difference between the conditions was confirmed to be statistically significant by repeating the test 10 times with different sets of faces from the database.

penalty term biased towards horizontal feature matches and third, the change in feature information for the same angular distance for faces is higher for vertical than for horizontal rotations.

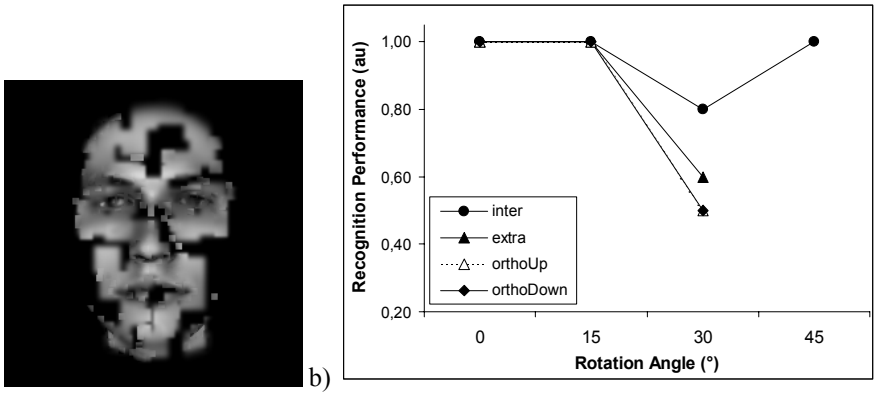

Figure 4. a) Feature representation as used by the computational framework – note that features focus on areas of interest (eyes, mouth, nose), b) machine recognition performance (arbitr. units) in the four rotation conditions (inter, extra, ortho up, ortho down) across viewpoint.

In Figure 4b, inter and extra conditions are plotted only for view-changes up to 30°. For larger view changes a global correspondence cannot be established anymore since the available feature sets are too different. This observation agrees with findings from a study in [27]. In order to address the issue of generalization over larger view changes, we propose the following extension to the framework, which consists of a two-step matching process.

First, in order to determine head position the image is matched on a coarse scale against different views from a database of faces. This is possible since the global facial feature layout guarantees a good pose recovery even for novel faces [30]. The second step then consists of using more detailed part layout information to match parts consisting of groups of features to the image in the corresponding pose. Parts (which would correspond in the ideal case to facial parts such as eyes, mouth, nose, etc.) can again be matched using the same algorithm as outlined above under the constraint of the global part layout information. Such a constraint can easily be built into the matching process as a prior on the allowed feature deformations. Again, this proposed framework is consistent with evidence from psychophysical studies (e.g., [13,21,22], see also [18] for a holistic two-stage model with alignment and view-interpolation).

In computational vision[4] the question how (facial) parts can be extracted from images and how a perceptually reasonable clustering of features can be created has recently begun to be addressed. A purely bottom-up way of extracting parts was suggested in [12], whereas [26,31] approach the issue from the perspective of categorization: extracting salient features, which maximize within-class similarity

[4] There is evidence from developmental studies that the basic schema of faces is innate [9,17], which could help newborn infants to learn encoding the parts of a face.

while minimizing between-class similarity. In [8], a 'Chorus of Fragments' was introduced, which is modeled after what/where receptive fields and also takes into account parts and their relations. One advantage of the framework proposed here, is its explicit use of features and their properties (such as pixel neighborhood, trajectory information, etc.), which provides the system with a rich representation and can be exploited for feature grouping. As shown in Figure 4a, the visual features already tend to cluster around facial parts and in addition also capture small texture features of the skin (such as birthmarks and blemishes), which were hypothesized [27] to be important features for less view-dependent face recognition.

General Discussion

Several previous studies have investigated face processing under varying pose (for a short review and further results see [24]). In order to further understand the viewpoint dependent nature of face recognition we investigated whether qualitatively similar effects of viewpoint apply to face recognition as found in studies using unfamiliar objects like wire-frames and amoeboid objects [5,7]. Indeed, this was the case, we found the same qualitative effects of viewpoint, which were consistent with a view interpolation model of object recognition [5,18,19]. In addition, a computational model based on local features and their relations [29] showed the same qualitative behavior as humans. The breakdown of this model for large view-changes motivates an extension of the framework to explicitly model parts. At the same time, this framework should provide greater robustness against partial occlusion and less susceptibility to viewpoint changes due to the use of parts [10,27].

The concept of representing objects by their parts and spatial relations has been proposed many years ago by structural description theories (e.g., [1,16]). There are, however, several important differences between these approaches and the framework we propose here. First of all, in contrast to the traditional approaches by Marr and Biederman, we are not convinced that it is biologically plausible and computationally possible to extract good edge-based representations as the input for recognition. Moreover, the parts we propose are completely different both conceptually and computationally from the geons used in the approaches in [1,11]. Geons are defined by using Lowe's nonaccidental properties [14] and are meant to be viewpoint-independent (or at least for a certain range of views, see [2]). In contrast to [3], we propose that face–recognition relies on processing local part-based and configural information (which could also apply to many cases of object recognition). In contrast to geons, the parts we propose are defined by grouping view-dependent image features. According to the type of features used, such parts are more or less viewpoint-dependent. We are currently running experiments, in which we explore to what extent *part-based* representations in human face recognition are viewpoint dependent. RBC theory assumes that a small set of geons suffices to explain the relevant aspects of human object recognition. In our view, in many cases of everyday object recognition, defining the features is a matter of perceptual learning [23], and

we believe that the number of parts represented by the human brain for recognition exceeds a 24 or 36 geon set by far and in addition might be heavily task-dependent.

References

1. Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. Psychological Review, 94(2), 115-147.
2. Biederman, I., Gerhardstein, P.C. (1993). Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. Journal of Experimental Psychology: Human Perception and Performance, 19, 6, 1162-1182.
3. Biederman, I., Kalocsai, P. (1997). Neurocomputational bases of object and face recognition. Philosophical Transactions of the Royal Society London, B, 352, 1203-1219.
4. Blanz, V. , Vetter, T. (1999). A Morphable Model for the Synthesis of 3D Faces. In Proc. Siggraph99, pp. 187-194.
5. Bülthoff, H.H., Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. PNAS USA, 89, 60-64.
6. Collishaw, S.M., Hole G.J. (2000). Featural and configurational processes in the recognition of faces of different familiarity. Perception, 29, 893-910.
7. Edelman, S., Bülthoff, H, H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research, 32(12), 2385-4000.
8. Edelman, S. Intrator, N. (2000). A productive, systematic framework for the representation of visual structure. In Proc. NIPS 2000, 10-16.
9. Goren, C., Sarty, M., Wu, P. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics, 56, 544-549.
10. Heisele, B., Serre, T., Pontil, M., Vetter, T., and Poggio, T. (2001). Categorization by learning and combining object parts. In Proc. NIPS 2001.
11. Hummel, J.E., Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. Psychological Review, 99(3), 480-517.
12. Lee, D., Seung S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401:788-791.
13. Leder, H., Candrian, G., Huber, O., Bruce, V. (2001). Configural features in the context of upright and inverted faces. Perception, 30, 73-83.
14. Lowe, D.G. (1985). Perceptual organization and visual recognition. Boston: Kluwer Academic Publishing.
15. Lowe, D.G. (1987). Three-dimensional object recognition from single two-dimensional images. Artificial Intelligence, 31, 355-395.
16. Marr, D. (1982). Vision. San Francisco: Freeman.
17. Morton, J., Johnson, M.H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. Psychological Review, 98, 164-181.
18. O'Toole, A. J., Edelman, S., Bülthoff H.H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. Vision Research, 38 , 2351- 2363.
19. Poggio T, Edelman S. (1990) A network that learns to recognize three-dimensional objects. Nature, 18, 343(6255), 263-266.
20. Pilu, M. (1997). A direct method for stereo correspondence based on singular value decomposition, In Proc. CVPR'97, 261-266.
21. Schwaninger, A., Mast, F. (1999). Why is face recognition so orientation-sensitive? Psychophysical evidence for an integrative model. Perception, 28 (Suppl.), 116.
22. Sergent J. (1985). Influence of task and input factors on hemispheric involvement in face processing. Journal of Experimental Psychology: Human Perception and Performance, 11(6), 846-61.

23. Schyns, P. G., Rodet, L. (1997) Categorization creates functional features. Journal of Experimental Psychology: Learning, Memory and Cognition, 23, 681-696.
24. Troje, N. F., Bülthoff, H.H. (1996). Face recognition under varying pose: the role of texture and shape. Vision Research, 36, 1761-1771.
25. Ullman, S., Basri, R. (1991). Recognition by linear combinations of models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(10), 992-1006.
26. Ullman, S., Sali, E. (2000). Object Classification Using a Fragment-Based Representation. In Proc. BMCV'00, 73-87.
27. Valentin, D., Abdi, H., Edelman, B. (1999). From rotation to disfiguration: Testing a dual-strategy model for recognition of faces across view angles. Perception, 28, 817-824.
28. Wallis, G. M. , Bülthoff, H.H. (2001). Effect of temporal association on recognition memory. PNAS USA, 98, 4800-4804.
29. Wallraven, C., Bülthoff, H.H. (2001). Automatic acquisition of exemplar-based representations for recognition from image sequences. CVPR 2001 - Workshop on Models vs. Exemplars.
30. Wallraven, C., Bülthoff, H.H. (2001). View-based recognition under illumination changes using local features. CVPR 2001 - Workshop on Identifying Objects Across Variations in Lighting: Psychophysics and Computation.
31. Weber M., Welling M. and Perona P. (2000). Unsupervised Learning of Models for Recognition. In Proc. ECCV2000.