Hidden Markov Model-based face recognition using selective attention

A.A. Salah[a], M. Bicego[b], L. Akarun[a], E. Grosso[b], M. Tistarelli[c],

[a]PILAB - Boğaziçi University, 34342 Bebek, İstanbul (Turkey)
[b]DEIR - University of Sassari, via Torre Tonda, 34 - 07100 Sassari (Italy)
[c]DAP - University of Sassari, piazza Duomo, 6 - 07041 Alghero (Italy)

ABSTRACT

Sequential methods for face recognition rely on the analysis of local facial features in a sequential manner, typically with a raster scan. However, the distribution of discriminative information is not unifom over the facial surface. For instance, the eyes and the mouth are more informative than the cheek. We propose an extension to the sequential approach, where we take into account local feature saliency, and replace the raster scan with a guided scan that mimicks the scanpath of the human eye. The selective attention mechanism that guides the human eye operates by coarsely detecting salient locations, and directing more resources (the fovea) at interesting or informative parts. We simulate this idea by employing a computationally cheap saliency scheme, based on Gabor wavelet filters. Hidden Markov models are used for classification, and the observations, i.e. features obtained with the simulation of the scanpath, are modeled with Gaussian distributions at each state of the model. We show that by visiting important locations first, our method is able to reach high accuracy with much shorter feature sequences. We compare several features in observation sequences, among which DCT coefficients result in the highest accuracy.

Keywords: Sequential face recognition, selective attention, saliency, scanpath, Gabor wavelets, HMM, DCT

1. INTRODUCTION

In recent years, a large number of methods have been investigated for automatic face recognition.[1] Among others, the sequential approaches received great interest.[2,3,4,5] These mimick the human visual system in its serial way of recognizing faces,[6] where the face image is explored with a scanning strategy, called a *scanpath*, in order to collect a sequence of features. For modeling the sequential data, Hidden Markov Models[7] have shown to be accurate and effective.[2,4,5] The main problem of this class of HMM-based approaches is that the scanning methodology is fixed (typically a raster scan is employed), and the importance (*saliency*) of facial parts is ignored.

This paper proposes a novel HMM-based face recognition system, in which a scanning strategy is employed to simulate a human-like saccadic sequence,[3] computed on the basis of the concept of saliency.[8] The approach converts a face image into an attention based "scanpath," that is, a sequence composed of two types of information: *Where* information, the coordinates of the salient region in the face, and *What* information, local features detected in there. At the core of the scanning mechanism is the calculation of saliency. This calculation should be cheap enough that it can be applied to the whole image without significantly increasing time and space requirements, and it should be informative. With this approach, a cheap and parallel search for salient features will drive a serial and detailed analysis.[9]

To generate the saliency map, we employ Gabor wavelets in multiple orientations and scales. The output of the scan-path generator mechanism is represented by a sequence of locations ("where" information, in the form of x,y coordinates) together with a sequence of sub-windows, from which different features could be extracted (we

Further author information: (Send correspondence to A.A.S.)
A.A.S.: salah@boun.edu.tr
M.B.: bicego@uniss.it
L.A.: akarun@boun.edu.tr
E.G.: grosso@uniss.it
M.T.: tista@uniss.it

compare graylevels as a baseline technique, against wavelet features and DCT-based features). These sequences are modelled using continuous HMMs, resulting in a standard Bayesian scheme.[10] The realized system has been tested on a subset of the BANCA database,[11] which contains faces from 52 subjects, gathered in 12 different sessions. The proposed algorithm is compared with raster scan-based methods using the same features. Our results show that the saliency-based system performs as well as the raster scan, even when a reduced number of saccades are used, leading to faster recognition. Our results were obtained using only the "what" information. A potentially useful combination of "where" and "what" information[3] is left as a future work.

The remainder of the paper is organized as follows: in Section 2, the fundamentals of the proposed approach, namely the HMM and the attention based scanpath scheme, are presented. Then, HMM-based face recognition systems are described in Section 3. The experimental evaluation is given in Section 4, followed by our conclusions in Section 5.

2. FUNDAMENTALS

This section contains the fundamentals: First, we review basic concepts regarding HMMs, then we present the attention based sequential scheme.

2.1. Hidden Markov Model

A discrete-time Hidden Markov Model $\boldsymbol{\lambda}$ can be viewed as a Markov model whose states cannot be explicitly observed: Each state has an associated probability distribution function, modeling the probability of emitting symbols from that state. More formally, a HMM is defined by the following entities:[7]

- $S = \{S_1, S_2, \cdots, S_N\}$ a finite set of hidden states;
- the transition matrix $\mathbf{A} = \{a_{ij},\ 1 \leq j \leq N\}$ representing the probability of going from state S_i to state S_j,
$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i]\ 1 \leq i, j \leq N$$
with $a_{ij} \geq 0$ e $\sum_{j=1}^{N} a_{ij} = 1$;
- the emission parameters $\mathbf{B} = \{b(o|S_j)\}$, indicating the probability of emission of the symbol o when the system state is S_j. In this paper we employ continuous HMMs: $b(o|S_j)$ is represented by a Gaussian distribution, *i.e.*
$$b(o|S_j) = \mathcal{N}(o|\mu_j, \Sigma_j)\,. \tag{1}$$
where $\mathcal{N}(o|\mu, \Sigma)$ denotes a Gaussian density of mean μ and covariance Σ, evaluated at o;
- $\boldsymbol{\pi} = \{\pi_i\}$, the initial state probability distribution, representing probabilities of initial states, i.e.
$$\pi_i = P[q_1 = S_i]\ 1 \leq i \leq N$$
with $\pi_i \geq 0$ and $\sum_{i=1}^{N} \pi_i = 1$.

For convenience, we denote an HMM as a triplet $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

The training of the model, given a set of sequences $\{\mathbf{O}_i\}$, is usually performed using the standard Baum-Welch re-estimation,[7] which determines the parameters $(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ that maximize the probability $P(\{\mathbf{O}_i\}|\boldsymbol{\lambda})$. In this paper, the training procedure is stopped after the convergence of the likelihood. The evaluation step, *i.e.* the computation of the probability $P(\mathbf{O}|\boldsymbol{\lambda})$, given a model $\boldsymbol{\lambda}$ and a new observation sequence $\mathbf{O}$, is performed using the *forward-backward procedure.*[7]

2.2. Attention-based sequential scan path extraction

The idea in selective attention based pattern recognition is to employ a computationally cheap screening process to select salient locations in an image. Historically, the models of human selective attention mechanism employ a large number of feature maps that evaluate saliency in different dimensions of the perceptual experience.[12] These features are then integrated to form a master saliency map that guides search in the image.[13] The location selection based on the saliency map is frequently performed with a Winner-Take-All (WTA) network, with which a natural segmentation of salient objects can be achieved.[9]

In this work, our aim is neither an accurate simulation of human eye movement behavior, nor the specification of the number and nature of feature maps that contribute to this behavior. We just want to implement a faster (and hopefully more accurate) face recognition system. To this end, we employ a simple saliency scheme. We use only bottom-up components, without any learned modulation function. We forgo the WTA in favor of a simple inhibition of return scheme. Fig. 1 depicts a schematic outline of the saliency model.

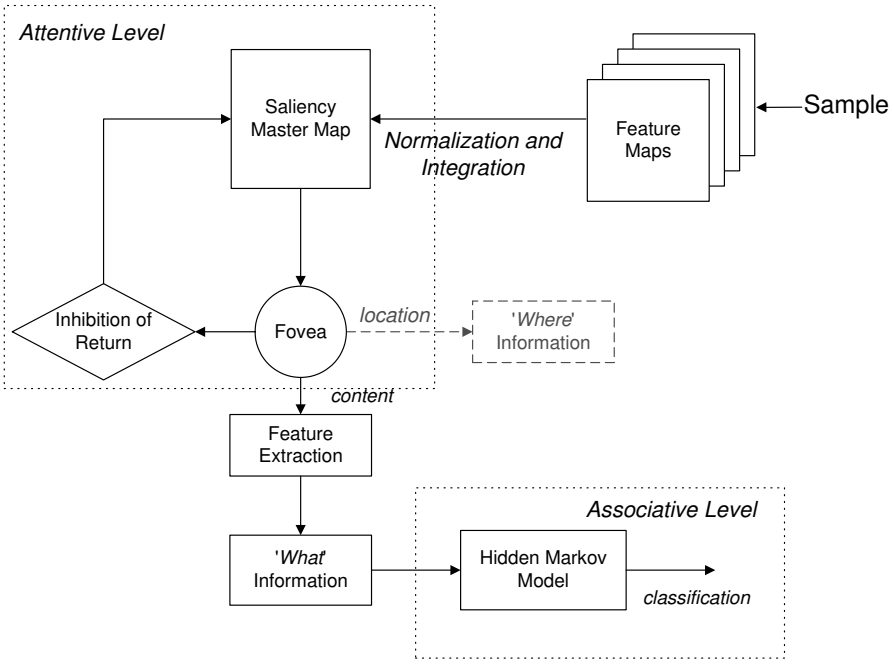

Figure 1. The saliency model.

We have used orientation sensitive Gabor wavelet functions in the construction of our feature maps.[14] The Gabor convolution can be expressed as:

$$\Psi_j(\vec{x}) = \frac{\vec{k}_j \vec{k}_j^T}{\sigma^2} e^{\left(-\frac{\vec{k}_j \vec{k}_j^T \vec{x}\vec{x}^T}{2\sigma^2}\right)} \left[e^{(i\vec{k}_j\vec{x})} - e^{(\frac{-\sigma^2}{2})}\right] \tag{2}$$

$$\vec{k}_j = (k_{jx}, k_{jy}) = (k_v \cos\varphi_w, k_v \sin\varphi_w), \quad k_v = 2^{-\frac{v+2}{2}}\pi, \quad \varphi_w = w\frac{\pi}{8}$$

where $\vec{x} = (x, y)$ indicates a pixel position in the image, j is the kernel index, (w, v) stand for the orientation and the scale parameters of the Gabor wavelet, respectively. The wavelets are evaluated on a downsampled version of the input, and eight Gabor maps with different parameter settings are selected experimentally. The responses are energy-normalized before integration, which is achieved via simple summation. The resulting saliency master map imposes an ordering on the image locations. The total saliency of the (overlapping) windows evaluated for the raster scan are computed for this purpose. The inhibition of return is permanent, as we never visit a window twice. This is not the case in biological systems, but we should note that biological systems always deal with active scenes, whereas we have static images.

3. FACE RECOGNITION USING HMMS

Face recognition using HMMs[2,4,5,15] typically involves the solution of different problems:

- **Coding**: A sequence of features is extracted from each face image. The ability of the system to discriminate among different faces strongly depends on the nature of extracted features, and their success in coping with the experimental conditions.
- **Learning**: The learning problem is solved by training a statistical classifier. Depending on the classification strategy chosen, it could be realized in different ways:
  1. Training one HMM for each class, subsequently using the standard Bayesian scheme for classification:[2,5] When a novel sequence is given, the posterior probabilities are calculated under each model, and the sequence is assigned to the class whose model shows the highest posterior probability.
  2. Training one HMM for each sequence: This scheme trains more models per class, provided that there is more than one training sample available per class. However, the amount of data used in training each model is smaller. This type of modeling approach can be said to compute distances between sequences; the classification in this case becomes a nearest neighbor rule in this likelihood sense.
  3. Mixed classification: There exist models that explore trade-offs between the two approaches we have mentioned. One can train one HMM for each sequence and for each class, making the classification scheme more complex, but also more powerful. Such a mixed approach is exemplified in Ref. 15, which is based on a dissimilarity-based representation.

Since the coding is crucial for the subsequent learning, we should take a moment to elaborate on it further. The strategy used to obtain the data sequence from a face image consists of two steps: scanning and feature extraction, respectively. In the the first step, a sequence of sub-images of fixed dimensionality is obtained by inspecting the face image. In the basic approach, that sequence is acquired by using a raster scan, i.e. by sliding a fixed sized window with a predefined overlap over the face image. The raster scan procedure is visualized in Fig. 2. The raster scan has some advantages, e.g. it is simple and exhaustive, but it also presents several

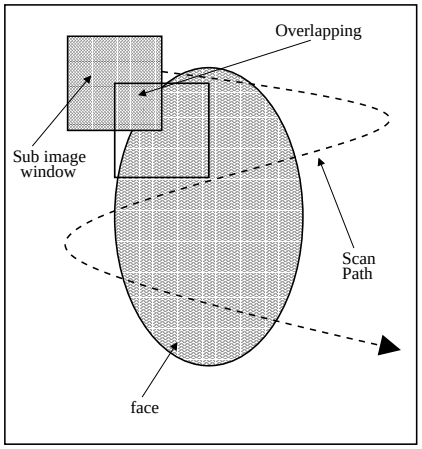

Figure 2. Sampling scheme to generate the sequence of sub-images.

problems:

- It requires registered images, if we want sequences obtained from different images to be on a par.
- It requires images with the background adequately removed. Otherwise, irrelevant and misleading data patches will periodically appear in the sequence.

- The analysis is blind, i.e. raster scan does not take into account that different parts of the face could convey different amounts of information.
- The whole image should be processed. This is both a blessing and a curse; the computational requirement is increased, but we are sure not to miss anything in the image.

In this paper, we seek a better scanning strategy, free of these handicaps. Our approach is the selective attention based scanning, presented in Section 2.2. The advantage of employing a saliency-based scheme is twofold: first, we can concentrate on the salient parts of the image, gaining robustness with respect to registration; second, since the patches are extracted in decreasing order of importance, we can decide to stop the analysis after a sufficient number of saccades. We will show in the experimental part that this could improve the classification accuracy of the classifier, eliminating the negative effect of misleading background frames to some extent. From the two types of information extracted by the saliency scheme, only the "what" information is used, leaving out the "where" features. The "where" information can potentially be useful to the analysis as well;[3] the integration of this type of information is left as a future work.

Once the sequence is extracted, the second step consists of computing features in each gathered sub-image. Statistical features, based on image moments or gradients, were previously employed for this purpose.[5] In Ref. 3, local neural network experts were employed to evaluate the sub-image, producing a class-posterior probability vector for each spot in the sequence that was used as the observation feature in the subsequent Markov model. This method provides a good leverage for the classification module, yet it is costly to maintain a large number of neural network experts. A more straightforward approach is to compute a few coefficients based on compression transformations. This is what we did in this paper, and several feature types were tested:

- Graylevel image intensity features are used as a baseline technique.
- DCT coefficients are extracted from each sub-image.[2] The obtained coefficients are scanned in a zig-zag fashion, similar to the method used for JPEG coding. Only few of these coefficients are retained, determining the dimensionality of the observation vector.
- Wavelet features are computed by substituting the DCT coding described above with the wavelet coding.[4] Haar wavelets[16] were used, representing the simplest wavelet basis. By using non-standard decomposition, which alternates between row and column processing, we achieve a more efficient coefficient computation. The proposed algorithm calculates the coefficients representing the image with a normalized two-dimensional Haar basis, sorting these coefficients in order of decreasing magnitude. Subsequently, the first M coefficients are retained, performing a lossy image compression. For a more complete treatment of wavelet image compression, see Ref. 16. As in the case of DCT, the number of retained coefficients determines the dimensionality of the observation vector.

By applying this step to all the sub-images of the sequence, we get the actual sequence observation. Its dimensionality will be $D \times T$, where D is the number of the features extracted from each sub-image, and T is the number of sub-images gathered in the sample scanning operation.

In this paper, the learning problem is addressed in a standard Bayesian manner:[10] one HMM is trained for each subject, and an unknown face (and relative sequence) is assigned to the class whose model shows the highest likelihood.

4. EXPERIMENTAL EVALUATION

The proposed methodology was tested using the BANCA database.[11] The part used for face authentication contains 52 subjects (26 female and 26 male). For each subject, 12 different sessions, recorded under different conditions are available (4 controlled, 4 degraded, and 4 adverse). We have selected to work on landmarked and registered images, to minimize the effect of preprocessing on reported recognition accuracy. In particular, all the images were preprocessed using a simple geometric normalization, followed by standard histogram equalization.[17] Geometric normalization maps each face on to a 55 pixel high by 50 pixel wide output image, via an affine transform that makes use of the manually annotated eye positions.

The protocol defined in[11] is designed for authentication, not for recognition. Therefore, we have used a Leave One Out (LOO) methodology[18] in this paper: each test sequence is removed from the set of available samples, a system is trained with remaining samples, and the test sequence is classified under the trained system. All possible test sequences are evaluated separately, and the results are averaged. We used only the first session of the database.

In this experimental scenario we performed different experiments, with different goals. In our first experiment the aim was to compare different feature extraction methodologies, to estimate the best parameters and the best strategy. In particular, we compare the following feature extraction methods:

- Gray levels of each sub-window
- DCT coefficients for a different number of retained coefficients
- Wavelets coefficients for a different number of retained coefficients

We have also tested various dimensions for the feature sub-window, whereas the overlap ratio was fixed at 0.5. Averaged LOO accuracies are displayed in Table 1. We report only the best results for DCT and Haar wavelet methods. The best results are obtained by using DCT features (only 4 coefficients are retained). Consequently, we will only report results with the DCT features in subsequent experiments.

Window Size	Methodology	LOO Accuracy
7	Gray levels	12.31%
7	DCT	85.00%
7	Haar wavelets	75.38%
9	Gray levels	17.69%
9	DCT	91.92%
9	Haar wavelets	59.62%
11	Gray levels	12.31%
11	DCT	94.62%
11	Haar wavelets	73.85%

Table 1. Evaluation of different feature extraction methods

The results of the first experiment suggest that the effects of the sub-window size should be explored more thoroughly. In order to have more statistically significant results (note that the HMM training which does not always converge to the same model), we have repeated the experiments of the first round five times, each time retaining four DCT coefficients, and averaged the results. Moreover in this experiment, we have compared the raster and saliency-based scanning methodologies. Note that when comparing these two methods, the content of the feature windows stays the same in each case, but the order changes. This is relevant for the HMM, which uses state transitions to model the temporal development of the sequence. Our results are displayed in Table 2.

Window Size	Average Acc. (std)		Max Acc.	
	Saliency	Raster	Saliency	Raster
7	87.62%(2.28%)	91.92%(1.63%)	91.15%	93.08%
9	89.31%(1.20%)	93.92%(0.92%)	90.38%	95.00%
11	93.69%(1.58%)	94.46%(1.29%)	95.77%	95.77%
13	95.23%(0.89%)	96.08%(0.74%)	96.15%	97.31%
15	96.85%(1.00%)	95.85%(1.29%)	98.08%	97.31%
17	93.15%(1.13%)	96.69%(0.89%)	95.00%	98.08%

Table 2. Comparison between raster and saliency-based scanning

With registered images, a raster scan gives a consistent sequencing for each image. Therefore, the presented results are conceivable as an upper-bound for the raster scan performance. The saliency-based scheme does not

make use of the alignment of the face images with respect to each other. Consequently, it is significant that the saliency-based scheme can reach a comparable accuracy without the use of registration information. In some of the experiments, the accuracy even exceeds that of the raster scan; indeed the best average accuracy is obtained with a saliency-based scheme.

The goal of the third experiment was to test the method based on saliency for a reduced number of saccades. Since the saliency imposes a natural ordering on the sub-windows, we can stop the algorithm before examining all the image. We would like to see how much accuracy is sacrificed by stopping earlier, and whether the computational savings would be worthwhile. The raster scan result is also reported for each experiment, which is however given after scanning the whole face. Results are displayed in Figure 3, for different sizes of windows. In general, the accuracy of the raster scan is reached by using only 40% of the saccades. This is a saving of 60% of computational time. Another benefit of the early stopping is the elimination of influences from less important or misleading image patches. Thus, we have cases where doing further saccades reduced the accuracy of the classifier.

5. CONCLUSIONS

We have presented a novel method for sequential face recognition, where a simple saliency-based scanpath simulation guides feature extraction. We use Gabor wavelets in saliency computation, and DCT compression features for the analysis of local patches. Hidden Markov models with Gaussian observation distributions are used for classification.

Our results indicate that the advantage of employing a saliency-scheme over a raster scan is twofold. By associating a measure of informativeness with image patches, we are able to devote more resources to discriminative locations, and discard clutter. Thus, we are able to save time in the classification phase by inspecting less than half of the image in detail. Furthermore, the saliency computation makes the algorithm robust with respect to registration, as the feature sampling follows salient locations instead of fixed points on the image plane.

Several directions are proposed for future work. Different saliency schemes will be tested to explore the trade-off between accuracy and computational complexity. The "where" information, i.e. the locations of the saccades can be integrated into the recognition algorithm. If the saliency scheme produces very similar scanpaths for each face, there will be no discriminative information in the "where" sequence. However, the local features of faces show great variation, and we expect that the "where" information will be useful too. Finally, we have reported our results in a small database. For more reliable results, we propose to employ larger databases with more difficult experimental settings.

6. ACKNOWLEDGEMENTS

This work is supported by EU FP6 Network of Excellence Biosecure.

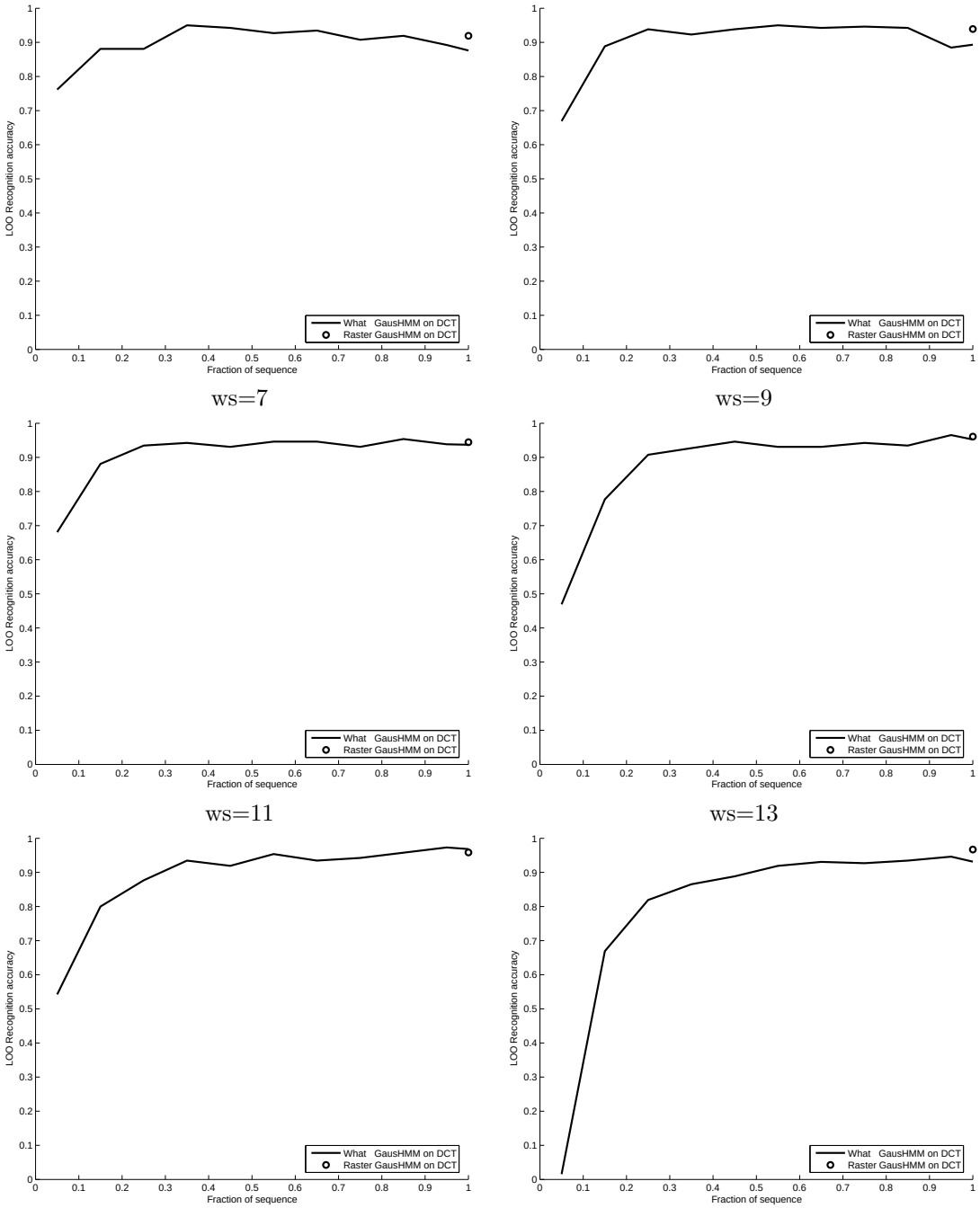

Figure 3. Leave-one-out accuracies for increasing lengths of sequences in the saliency-based approach. The full-sequence raster scan accuracy is also shown as a single point. Roughly 40 per cent of the sequence is sufficient to reach a comparable result with the raster scan.

REFERENCES

1. W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys* **35**, pp. 399–458, 2003.
2. V. Kohir and U. Desai, "Face recognition using DCT-HMM approach," in *Proc. Workshop on Advances in Facial Image Analysis and Recogniti Technology (AFIART)*, (Freiburg, Germany), 1998.
3. A.A. Salah, E. Alpaydın, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24**(3), pp. 420–425, 2002.
4. M. Bicego, U. Castellani, and V. Murino, "Using Hidden Markov Models and wavelets for face recognition," in *IEEE. Proc. of Int. Conf on Image Analysis and Processing*, pp. 52–56, 2003.
5. M. Bicego, E. Grosso, and M. Tistarelli, "Probabilistic face authentication using hidden markov models," in *Proc. of SPIE Int. Workshop on Biometric Technology for Human Identification*, 2005.
6. D. Noton and L. Stark, "Eye movements and visual perception," *Scientific Ann.* **224**, pp. 34–43, 1971.
7. L. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. of IEEE* **77**(2), pp. 257–286, 1989.
8. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20**(11), pp. 1254–1259, 1998.
9. C. Koch and S. Ullman, "Shifts in selective visual-attention towards the underlying neural circuitry," *Human Neurobiology* **4**, pp. 219–227, 1985.
10. R. Duda, P. Hart, and D. Stork, *Pattern Classification*, John Wiley and Sons, 2^{nd} ed., 2001.
11. E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA database and evaluation protocol," in *Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA03)*, pp. 625–638, Springer-Verlag, 2003.
12. A.M. Treisman and G. Gelade, "A Feature Integration Theory of Attention," *Cognitive Pscychology*, **12**(1), pp. 97–136, Jan. 1980.
13. L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, **2**(3), pp. 194–203, Mar. 2001.
14. J.G. Daugman, "Complete Discrete 2D Gabor Transform by Neural Networks for Image Analysis and Compression," *IEEE Trans. Acoustics, Speech, and Signal Processing*, **36**, pp.169–179, July 1988.
15. M. Bicego, V. Murino, and M. Figueiredo, "Similarity-based classification of sequences using hidden markov models," *Pattern Recognition* **37**(12), pp. 2281–2291, 2004.
16. R. D. Vore, B. Jawerth, and B. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. on Information Theory* **38**(2), 1992.
17. R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, 2^{nd} ed., 2002.
18. S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.