Stochastic Models for Face Image Analysis

Stéphane Marchand-Maillet and Bernard Mérialdo

Department of Multimedia Communications
Institut EURECOM – B.P. 193
06904 Sophia-Antipolis – France
{marchand,merialdo}@eurecom.fr

Abstract. This study continues our work on using stochastic models for image analysis in the context of video indexing. Pseudo-two dimensional Hidden Markov Models (P2DHMM) were shown to be efficient and flexible tools for performing human face localisation in colour images from video sequences. In this context, little constraints can be applied on face images for their identification in view of indexing. Here, we present a technique based on P2DHMM for recovering face orientation in colour images cropped from video sequences. Such a procedure will ease identification by either direct comparison or clustering. Results are presented which show the accuracy of our technique and confirm the capabilities of such stochastic models in the task of video indexing.

1 Introduction

The indexing of video data calls for the analysis of its frames both in the spatial and temporal domains. While temporal analysis provides information on the dynamic features of the data, spatial analysis allows for the identification of objects and persons featuring within the sequence. By contrast with images used for classical recognition tasks such as face identification from mugshots, face images obtained from video sequence vary substantially in size, orientation and illumination. It is therefore important that models used in this context are flexible enough to cope with such a lack of constraints.

In this paper, we continue the work presented in [4] where stochastic models were designed for colour image analysis. It was shown in [4] that pseudo two-dimensional Hidden Markov Models (P2DHMM) could be used for achieving the goal of segmenting faces in images extracted from video sequences. We extend this work by presenting here a technique based on similar models which recovers the orientation of faces from images cropped from a video sequence. The aim of this step is to provide enough information for performing subsequent identification by clustering or direct comparison.

The paper is organised as follows. We first review in Section 2 the principles behind P2DHMM and the context in which these tools will be used. Then, Section 3 specialises in designing and training models for recovering face orientation. Section 4 details the results obtained from experiments we made using the technique presented in this paper. Finally, Section 5 gives some conclusions on this approach and suggests further extensions in the context of video indexing.

2 Stochastic modelling of images

In this section, we briefly recall the principles behind Hidden Markov Modelling, starting with well-known mono-dimensional models and extending to pseudo two-dimensional models.

2.1 1D Hidden Markov Models

Hidden Markov Models are powerful tools for recovering the structural information from a given observation sequence. 1DHMM are commonly used in speech processing where the observation sequence relates the evolution of speech parameters over time. For a detailed presentation of these tools, the reader is referred to [6,7].

Formally, a Hidden Markov Model (HMM) is a finite state machine composed of N states s_i, $i = 1, \ldots, N$, characterised by their associated output probability functions. Typically, each state s_i is able of outputting any observation value v with a probability $b_i(v)$. Transitions between states s_i and s_j are constrained by the transition probability a_{ij}. The matrix $A = (a_{ij})$ of all transition probability values therefore provides information on the topology of the HMM. Here, we use only left-right HMM, meaning that states are ordered and that a state s_i leads only either to itself (with the probability a_{ii}) or to the next state s_{i+1} (with the probability a_{ii+1}). The transition probability matrix will therefore contain non-zero values on its diagonal and upper-diagonal only. Initial state probabilities allow for constraining the state corresponding to the first observation in the sequence (*i.e.*, the state in which the HMM starts outputting the observation sequence). In our context we also force the HMM to terminate the process in its last state s_N (see Figure 1).

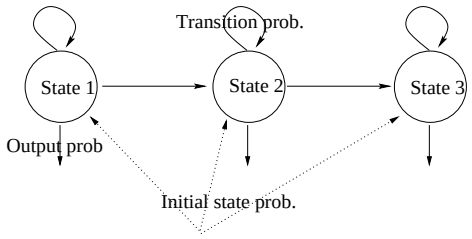

Fig. 1. A left-right (Bakis) Hidden Markov Model.

Given a 1D observation sequence $O = (o_1, \ldots, o_X)$ and parameters, let $Q = (q_1, \ldots, q_X)$ be a sequence of state associated with O (*i.e.*, $q_x = s_i$ for some i and for a given x). The likelihood of O being created through the state sequence Q by the HMM Λ is given by

$$P[O|Q, \Lambda] = \prod_{x=1}^{X} b_{q_x}(o_x),$$

and the probability of O given the model Λ is obtained by summing the joint probability of O and Q over all possible state sequences Q. In other words,

$$P[O|\Lambda] = \sum_{\text{all possible } Q} b_{q_1}(o_1)a_{q_1q_2}b_{q_2}(o_2)a_{q_2q_3}\ldots b_{q_X}(o_X)$$

Conversely, given O and parameters for the HMM Λ, using the Viterbi procedure based on dynamic programming, one can recover the most likely state sequence Q associated with O, thus recovering the (hidden) structure of O.

Finding optimal model parameters for describing a given class of observation sequences is done by training the HMM using representatives of this class. That is, given initial model parameters (usually found by statistics on hand-segmented sequences), each observation sequence O_k is associated with a novel state sequence Q_k. Using the statistics of these new segmentation patterns (Viterbi training) or re-estimation formulae (Baum-Welch training), the parameters of the model Λ are updated and the procedure iterated. This training process is shown to converge toward a set of model parameters maximising the goodness-of-fit of the model with respect to the training sequences ($P[O_k|\Lambda]$).

2.2 Pseudo-2D Hidden Markov Models

In [4], it was shown that modelling an image using a 1DHMM for each type of line was possible but that these models where lacking coherence in their perpendicular direction (vertical direction). On the other hand, it is known that the extension of the Hidden Markov Model to the 2D case leads to an unmanageable computational complexity. In this respect, we detailed in [3,4] the use of *pseudo* two-dimensional HMM (P2DHMM) for colour image analysis. Extending the technique first presented in [1,2], we could successfully segment faces from video images without imposing constraints on their orientation.

The idea of a P2DHMM is to consider a set of 2D-data as being composed of lines which can be accurately modelled by a set of N_M 1DHMM λ_i and to constrain the sequence in which these 1D models are fitted with the lines by a vertical 1DHMM Λ. In other words, the horizontal models λ_i become the (super-)states of the vertical model Λ. Vertical transitions can therefore only be defined at the line level, by opposition with the pixel level (see Figure 2). Output probabilities of super-states are implicitly given by the goodness-of-fit of the corresponding horizontal model λ_i with respect to the line L_y in question (*i.e.*, $P[L_y|\lambda_i]$).

This extension is therefore fairly straightforward and practical implementation is done by considering a doubly-embedded training procedure where 1D horizontal HMM are trained and define parameters for the super-states of the vertical HMM which is trained in a similar way.

3 Face image analysis

We now use the above technique to model face images in a way that face orientation can be recovered from the image.

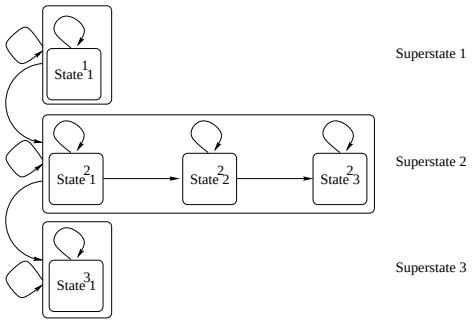

Fig. 2. A pseudo two-dimensional extension of the 1DHMM shown in Figure 1.

In [4], the model of an image containing a face was constructed in order to locate this face within the image. A P2DHMM was used where states were specifically defined as Background state and Face state. The segmentation was then simply done by considered the face region the set of pixels whose corresponding states in the best state sequence was Face. It was argued that adding complexity within the model was made difficult by the high flexibility of the models. In this example application, it was not possible to model states such as eyes, mouth or nose. The reason was a follows. As soon as a level of detail is set for the model then all details at this level are to be handled. In this application, the background being unknown, it was not possible to model such a level of detail. Following this analysis, we will use a P2DHMM for modelling a face image.

3.1 Face image models

We choose to model the complete image by a 7×7 state grid. That is, 7 superstates of 7 states each which will provide a 49-label partition of the face image. The image is partitioned into 8×8 blocks overlapping by half (4 lines and 4 columns). Following [5], the observation vector v at each pixel is composed of the 4 first components of the 2D-DCT of the block corresponding to that location.

We model the output probability function of each state as a mixture of 4D-Gaussian functions.

$$b_i(v) = \sum_{m=1}^{M} c_m \prod_{d=1}^{4} \left[\frac{1}{\sqrt{2\pi\sigma_m^{(d)}}} \exp\left(\frac{-\left(v^{(d)} - \mu_m^{(d)}\right)^2}{2\sigma_m^{(d)}} \right) \right]$$

The parameters to maintain for each output probability function are therefore the mixture coefficients c_m and the Gaussian function parameters (mean μ_m and variance σ_m).

3.2 Training and recovering face orientation

Training is done in an unsupervised way using the Baum-Welch training procedure. For each orientation, a set of training images is created. The observation

values are presented to the oriented face model along with corresponding initial state values. Since we do not aim to create any particular one-to-one mapping between states and face parts, the original state map is simply a regular 7×7 grid of states (*i.e.*, the image is partitioned in 7×7 regular regions, each of which is given a state number in sequential order). Initial state parameters are then found using LBG (k-Means) algorithm and the training procedure is iterated until convergence.

One model Λ_i is trained per orientation (angle θ_i) with instances of face images exhibiting this orientation. Recovering the orientation of an unknown face image I is done using the Viterbi algorithm for calculating $P[I|\Lambda_i]$, the goodness-of-fit of each model Λ_i against the image I. The model Λ_{i^*} which results in the maximal probability value is selected so that its index i^* indicates the orientation θ_{i^*} of the face under investigation.

4 Results

From the video AIM1MB08 of the reference video indexing database created by INA (Institut National de l'Audiovisuel, France), we created a set of 499 face sub-images cropped from scene where a character is moving (*e.g.*, rotating) his head, thus obtaining the face in different orientations. All these images were normalised to a size of size 64×64 and associated by a human operator with an index representing the value of the rotation angle and the sign of the rotation. Nine index values ($i = -4, \ldots 4$) were therefore obtained indicating the value of the (positive or negative) angle from front view $\theta_i = i * 22.5^\circ$ (0=front view, 4=profile view).

20 images were used for training each model Λ_i corresponding to a particular orientation θ_i. The models were trained until the relative improvement in the log-likelihood over all images was lower than 0.05%. This corresponds to running about 10 iterations using all images.

The tests were conducted over all images and results are presented for the test and the training images separately. Each image I was associated with 9 values $\{P[I|\Lambda_i] \mid i = -4, \ldots, 4\}$ and the orientation θ_{i^*} finally recovered for the image is such that

$$i^* = \arg \max_{i=-4,\ldots,4} P[I|\Lambda_i].$$

Figure 3 (right) shows the value of the model likelihood values obtained for the first 100 images of the test set.

First, we note the value of the log-likelihood at about -1.2×10^4. This low value is due to the size of the images added to the fact that the pseudo-2D combination of state transitions is represented by a product of probability values, thus taking a value far smaller than 1. This graph shows that for most images, the best likelihood value is well-separated form other values. On the other hand, for some instances, this is not the case, suggesting that more information may be obtained by exploiting also second best and other values rather than only considering the maximum value.

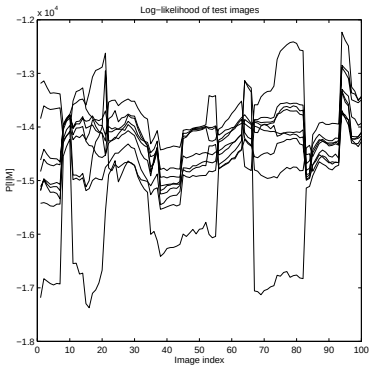

Fig. 3. Log-likelihood of the models against images.

Figure 4 (left) compares the value of the orientation obtained with the original value for the 319 test images. The step-like curve shows the evolution of the orientation given by a human operator (ground truth) through the test set and for each image, a dot gives the corresponding orientation found using our models.

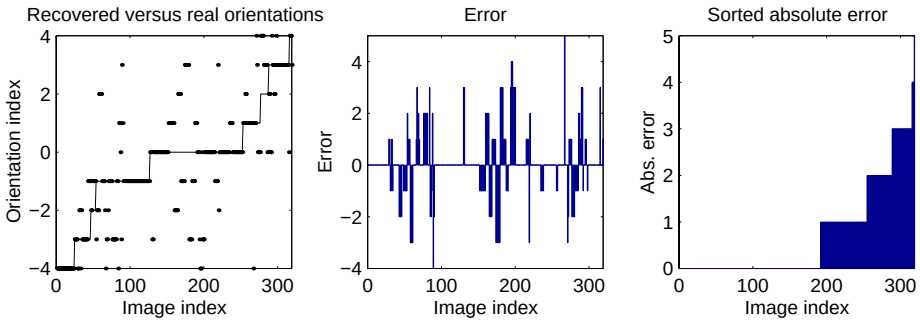

Fig. 4. Comparison between recovered and real face orientations.

On average, we found an error of ± 0.6794, indicating that an error of about $\pm 15.2°$ is made when recovering face orientation. Results were divided as follows. On the training images, the error value is of ± 0.6111 ($\pm 13.7°$) and for test images, this value is of ± 0.7179 ($\pm 16.15°$). Figures 4 (centre and right) detail these errors for the test set. Typically, over 319 images, 191 ($\simeq 60\%$) of the orientation values were recovered exactly and 63 images showed an absolute error of ± 1 point in the index (*i.e.*, $\pm 22.5°$ in the orientation). If we considered that an absolute error no larger than 1 point in the orientation is tolerated, we obtain about 80% as success rate.

These results should be modulated by the fact that ground truth data may not be perfect since subjectivity is present in their creation, thus making the tolerance of ± 1 point realistic. Moreover, we use face images cropped directly from video sequence without background removal. This surely affects the models likelihood values and using the technique presented in [4] for precise background removal may better the results. Nevertheless, results obtained are definitely promising.

Figure 5 shows a random selection of images sorted by their orientation found with our models. The value of the orientation index found is indicated above each picture and, between brackets, we show the error measured with the human-found orientation index.

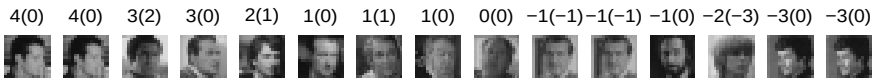

Fig. 5. Random sequence sorted with respect to face orientation.

Apart from the third image from the right which shows an error of -3 in its index (*i.e.*, $-45°$ instead of $+225.5°$), the sequence in which the images are sorted can be considered as fairly correct. Moreover, our training may be biased by the fact that only few different persons were present in the database so that a part of the model is influenced by the identities of the characters. In this context, since some characters did not appear for some specific orientations, the evaluation of their images with respect to models trained without instances of their faces have been altered by their identity. An improvement in these results may therefore lie in the enrichment of the training database for making the models more independent to the person identity.

5 Conclusion

In this paper, we presented a technique for recovering face orientation from images extracted from a video sequence. This study was developed in the context of video indexing where little or no constraints can be applied on face in terms of size, orientation and illumination. Continuing the work presented in [3, 4], we used pseudo two-dimensional Hidden Markov Modelling for achieving our aim.

Models were designed and trained, each of them corresponding to a particular face orientation. Our results showed that stochastic modelling could achieve the recovering of face orientation. These experiments need to be confirmed in a wider context, where models are trained in a database making them more independent to person identity.

On the other hand, the bias of the models may be exploited in the context of person identification. Further work using HMM in video indexing may therefore

include face recognition where HMM are used for recognition and the techniques presented in [5,8] combined with this technique for accurate face recognition.

Exploiting the temporal information contained within the video may also benefit the results when recovering face orientation. Using a global 1DHMM modelling the sequence in which orientation models are associated with face images, one may remove estimation errors by adding coherence within the orientation sequence. For example, when it is known that a continuous face rotation is shown in a particular sequence, this upper-level HMM can therefore be used for determining steps in the sequence between which face images can be considered as similar.

Acknowledgements

Eurecom's research is partially supported by its industrial partners: Ascom, Cegetel, France Telecom, Hitachi, IBM France, Motorola, Swisscom, Texas Instruments, and Thomson CSF.

References

1. O. Agazzi and S.-S. Kuo. Joint normalization and recognition of degraded documents images using pseudo-2D HMM. *Proceedings of the IEEE*, pages 155–158, 1993.
2. S.-S. Kuo and O. E. Agazzi. Keyword spotting in poorly printed documents using Pseudo 2-D Hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-16(8):842–848, 1994.
3. S. Marchand-Maillet. 1D and pseudo-2D Hidden Markov Models for image analysis. Technical Report RR-99-49, Institut EURECOM, Dept of Multimedia Communications, 1999.
4. S. Marchand-Maillet and B. Mérialdo. Pseudo two-dimensional Hidden Markov Models for face detection in colour images. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.
5. A. V. Nefian and M. H. Hayes III. Face recognition using an embedded HMM. In *Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington DC, USA, 1999.
6. L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, 1989.
7. L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.
8. F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceeding of the Second IEEE Workshop on Applications of Computer Vision*, Sarasota, Florida, December 1994.