Robust Head Gestures Recognition for Assistive Technology

Juan R. Terven[1], Joaquin Salas[1], and Bogdan Raducanu[2]

[1] Instituto Politécnico Nacional
Cerro Blanco 141, Colinas del Cimatario, Queretaro, Mexico
jrterven@hotmail.com,
jsalasr@ipn.mx
[2] Centre de Visió per Computador Edifici "O", Campus UAB
08193, Bellaterra, Cerdanyola (Barcelona) Spain
bogdan@cvc.uab.es

Abstract. This paper presents a system capable of recognizing six head gestures: nodding, shaking, turning right, turning left, looking up, and looking down. The main difference of our system compared to other methods is that the Hidden Markov Models presented in this paper, are fully connected and consider all possible states in any given order, providing the following advantages to the system: (1) allows unconstrained movement of the head and (2) it can be easily integrated into a wearable device (*e.g.* glasses, neck-hung devices), in which case it can robustly recognize gestures in the presence of ego-motion. Experimental results show that this approach outperforms common methods that use restricted HMMs for each gesture.

Keywords: assistive technology, social interactions, head gestures recognition, hidden markov models.

1 Introduction

During the last decades, several researchers and companies have developed assistive technology systems to support visually impaired people in tasks such as navigation, orientation, object recognition, and reading printed material [1]. However, few efforts have been devoted in systems to enhance their social interaction [2]. The issue is that during social interactions, a large part of the communication is non-verbal [3] and is given mostly in the form of visual cues such as head and body gestures. Moreover, most sighted people are not aware of the non-verbal signals that they commonly use and make no adjustment when interacting with visually impaired people [2]. In turn, this leaves visually impaired people in disadvantage, promoting in some cases social isolation [4].

Some of the requirements of social interaction devices, in the opinion of Krishna *et al.* [2], include the identification of the following: (1)number of participants, (2)gaze direction, (3)people's identity, (4)people's appearance (face, clothes), (5)facial expressions, and (6)hand and body gestures. Besides these,

J.F. Martínez-Trinidad et al. (Eds.): MCPR 2014, LNCS 8495, pp. 152–161, 2014.

in our research, we have identified the need of inferring the degree of attention that the interlocutor is paying to the speaker during social interactions. This need arises, for instance, when a visually impaired person is engaged in a dyadic interaction (one-to-one) and the interlocutor decides to walk away inadvertently.

In this paper, we develop a highly effective method to recognize head gestures, such as shakes and nods, as they represent an important part of the non-verbal communicative process and, together with the non-verbal backchannels, could signal agreement, disagreement, or the intention for turn-taking [5]. The rest of the paper is structured as follows. In Section 2, we survey previous work on the visual detection of head gestures. Then, in Section 3, we describe our methodology, based largely on the use of Hidden Markov Models (HMM). In section 4, we report experimental results on two datasets. Finally, we conclude and suggest future lines of work.

2 Related Work

In this section we survey previous work on head gesture recognition, highlighting the strengths and weakness for assistive technology applications. According to our review, most head gesture recognition methods divide the process in two steps: motion estimation and temporal sequence analysis.

For motion estimation, several authors have performed head gesture recognition based on eye tracking [6–9]. For instance, Choi & Rhee [6] and Kang *et al.* [7] segmented the eyes by thresholding, and making use of the head's geometrical characteristics; Kapoor *et al.* [8] used an infrared sensitive camera and infrared LEDs to track pupils; and Tan & Rong [9] used the cascade classifier from [10] to detect and track the eyes. Nevertheless, a disadvantage of using the eyes as features for motion estimation, is the inability to detect and track faces with sunglasses (commonly worn by visually impaired people). Gunes & Pantic [11] and Fujie *et al.* [12] estimated head motion using optical flow. More precisely, the head region was extracted by skin color segmentation. This approach works well in controlled environments. Yet, methods based on color are sensitive to changes in illumination and therefore, are not suitable for outdoors. In [11], the head region is detected using the cascade classifier from [10]. However, using [10] for tracking is computationally expensive for an embedded device and the face angles are confined to a limited range. Wei *et al.* [13] used the Kinect sensor to detect and track the head pose (yaw, pitch, roll). Although this sensor is low cost and provides great capabilities, its use is limited to indoors or dark environments and is not suitable for a wearable device. To overcome these limitations, we use a robust method for motion estimation based on SIFT features [14], capable of detecting and tracking the face even with sunglasses with various illuminations.

The second step of head gesture recognition is temporal sequence analysis. A common approach for this is to use HMMs to recognize each gesture. In fact, all the systems described before make use of two or three HMMs with different states [6–9, 11–13], *e.g.,* up/down for nodding and left/right for shaking. This approach works well if the system is in a fixed position (*i.e.,* third-party

perspective), and is commonly use in applications such as robotics and Human Computer Interaction (HCI), where the system should be able to infer the proper gestures and react accordingly. However, if the system is moving, for instance implemented in a wearable assistive device, it should be able to robustly identify the head gestures, despite the ego-motion noise. To account for this, we propose a more flexible approach using identical ergodic HMMs that include all the movements.

3 Methodology

In this section, we describe the process to compute the head pose, and infer nod and shake movements.

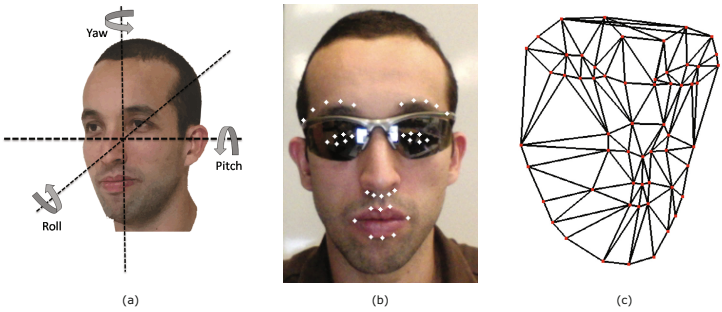

Fig. 1. Head Pose Estimation. (a) If we consider the head as a disembodied rigid object, its pose can be described by yaw, pitch, and roll (b) 2D Face features extracted using the Supervised Descent Method (SDM). (c) 3D anthropometric head model. This 3D model and the 2D features are fed into the POSIT algorithm[15] for head pose estimation.

3.1 Head Pose Estimation

Head pose estimation refers to compute the orientation of a person's head with respect to a camera [16]. If we consider the human head as a rigid object, the pose estimation is limited to three degrees of freedom characterized by *yaw*, *pitch*, and *roll* as shown in Fig. 1(a).

To estimate the head pose, we detect facial landmarks using the *Supervised Descent Method* (SDM) proposed in [17] and freely available online[1]. SDM is used for solving the Non-linear Least Squares (NLS) problem of face alignment. Given a set of n training images $\mathbf{d}^i \in \Re^{m\times 1}, 1 \leq i \leq n$, of m pixels with p landmarks. The system uses SIFT [14] features $\mathbf{h}(\mathbf{d}(x)) \in \Re^{128\times 1}$ extracted from patches around the landmarks (shown in Fig. 1(b)) to learn a series of descent directions

[1] Human Sensing Laboratory: `http://www.humansensing.cs.cmu.edu/intraface` (Accessed March 25, 2014)

and re-scaling factors such that it produces a sequence of updates in order to converge from the initial estimate (x_0) to the ground truth landmarks (x_*). With this approach, face alignment consists of minimizing the function over Δx

$$f(x_0 + \Delta x) = \|\mathbf{h}(\mathbf{d}(x_0 + \Delta x)) - \mathbf{h}(\mathbf{d}(x_*))\|_2^2 \,. \tag{1}$$

Once the face is aligned, *i.e.,* the landmarks converge to the final position, the head pose estimation is obtained following the approach described in [18], where 2D image points and 3D model points are matched, using the POSIT (*Pose from Orthography and Scaling with Iterations*) method [15]. The 3D anthopometric model (Fig. 1(c)) that we used is presented in [18] and it is available online[2].

3.2 Head Nodding and Shaking Detection Using Two HMMs

Once estimated the head pose, the next step is to recognize head gestures such as nodding and shaking. For this purpose, we started developing two HMMs shown in Fig. 2. One is used to recognize nodding, and the other to recognize shaking.

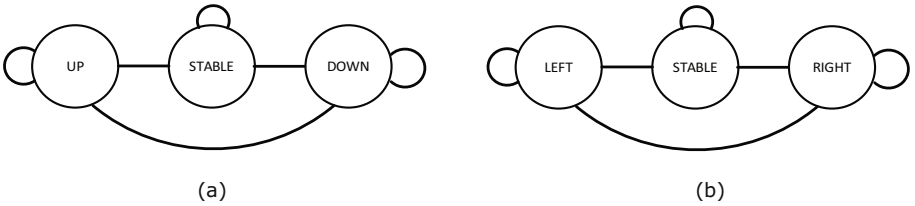

Fig. 2. Hidden Markov Models (HMM) designed for head gestures recognition. (a) HMM for nodding recognition, (b) HMM for shaking recognition.

Hidden Markov Models (HMMs). A Hidden Markov Model is a double embedded stochastic process: An underlying stochastic process that is not observable, and a set of stochastic processes that produce the sequence of observations [19]. Each HMM is defined by states, states transition probabilities, observation probabilities, and initial probabilities as follows:

1. The N states in the model $S = \{S_1, S_2, \ldots, S_N\}$. In our case both HMMs have three states: *stable*, *up*, and *down* for the nodding HMM; and *stable*, *left*, and *right* for the shaking HMM.
2. The M observation movements per state $V = \{v_1, v_2, \ldots, v_M\}$. In our system the movements are *stable*, *upward*, *downward*, *leftward*, and *rightward*.
3. The state transition matrix $A = \{a_{ij}\}$, where a_{ij} is the probability that the state S_j is given at time $t + 1$, when the state at time t is S_i.

[2] Advance Interaction using Facial Information: `http://aifi.isr.uc.pt/index.html` (Accessed March 25, 2014).

For example, let's consider $N = 3$, and S_1 is *stable*, S_2 is *up*, and S_3 is *down*. Therefore, a_{11} is the probability of transitioning from *stable* to *stable*. Similarly, a_{12} is the probability of transitioning from *stable* to *up*, and so on.

4. The observation probability matrix $B = \{b_j(k)\}$, where $b_j(k)$ is the probability that the symbol v_k is emitted in the state S_j.
5. The initial state distribution $\boldsymbol{\pi} = \{\pi_i\}$, where π_i is the probability that the model starts at state S_i. In our system we choose $S_1(stable)$ as the initial state.

In practice, the state transition matrix A and the observation probability matrix B are learned during training.

Training. For training and test we collected a database of 10 videos from different people, taken from a static webcam and 10 videos taken with the Pivothead[3] wearable glasses. All the videos contain annotated groundtruth and can be downloaded from the web[4]. Each gesture in the video is translated into a sequence or time series of 20 digits long containing the changes in yaw and pitch in consecutive frames. Fig. 3 shows typical nodding and shaking sequences from the database. From these graphs, we can see that a nodding gesture exhibits larger changes in pitch than in yaw. Conversely, a shaking gesture exhibits larger changes in yaw than in pitch. With these clearly distinctions, it is easy to extract simple movements from the time series.

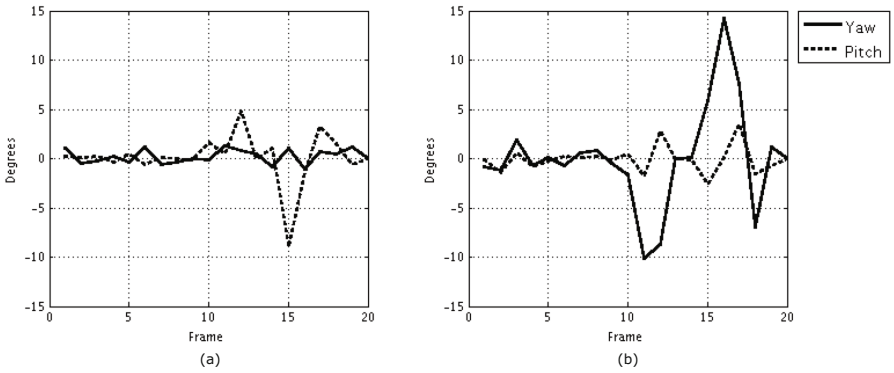

Fig. 3. Yaw and Pitch changes for typical nodding and shaking sequences. (a) shows the yaw and pitch changes (in degrees) for a nodding gesture. (b) shows the yaw and pitch changes (in degrees) for a shaking gesture.

As mentioned before, we defined five observation movements: *stable*, *upward*, *downward*, *leftward*, and *rightward*. Each move is represented by a number (from 1 to 5), and for each training time series, we extracted a sequence of these

[3] Pivothead Wearable Imaging: `http://pivothead.com` (Accessed March 25, 2014).
[4] Head Gestures Dataset: `www.jrterven.com/headDataset.html` (Accessed March 25, 2014).

movements from the yaw and pitch changes using the following procedure: Let Δx represent the change in yaw in two consecutive frames. Likewise, Δy represents the change in pitch in two consecutive frames. If $|\Delta y| \gg |\Delta x|$ then the symbol is *up* or *down* (*i.e.,* 2 or 3), now we look at the sign, if Δy is positive, the extracted symbol is 2 (*up*), but, if it is negative, the symbol is 3 (*down*). On the contrary, If $|\Delta x| \gg |\Delta y|$ then the symbol is *left* or *right* (*i.e.,* 4 or 5) and a similar procedure is followed for extracting the final symbol. If none of these conditions are true, the symbol is 1 (*stable*). After this procedure, we are left with a sequence of 20 digits. For example, the sequence (1 1 1 1 2 2 1 1 1 1 1 1 1 3 3 1 1 1 1 1) stands for an *stable* phase followed by *upward* then *stable* followed by *downward*, and the ground truth indicates that this sequence is a nod gesture. The set of all these sequences with their corresponding ground truth constitutes the training data for the HMMs. To select the optimal value for discriminating between vertical and horizontal movements (the $\gg$ threshold), we tested the algorithm with different discriminator thresholds and picked the one with the highest F-score.

The goal of the training phase is to estimate the state transition matrix (A) and the observation probability matrix (B). Using these sequences, we trained two HMMs using the Baum-Welch algorithm: one for nodding and one for shaking. However, the multiple repetitions of the state S_1(*stable*) produce a very high probability of transition from *stable* to *stable* (≈ 0.74 for nodding and 0.76 for shaking). These high probabilities affect the recognition, because movements containing multiple *stable* states (even a static head) can be regarded as nods or shakes with high probabilities. To address this problem, we removed the repetitions of the *stable* observations from the training and testing sequences. Now, instead of having sequences such as (1 1 1 1 2 2 1 1 1 1 1 1 1 3 3 1 1 1 1 1) for nodding, we are left with sequences like (1 2 2 1 3 3 1). This change improved the recognition stage because now each sequence is shorter and the recognition values are not affected by long chains of *stable* symbols. The repetitions of the other observations are necessary for recognition, because a single observation indicates a fast movement and repetitions of the same observation indicate a slow movement.

Recognition. Given an observation sequence extracted from video, the goal of recognition is to determine which one of the two HMMs is more likely to have generated the sequence. In our case, we used the Baum-Welch algorithm to obtain the probabilities of the observation sequence given each model. To determine the gesture, we selected the model with the highest probability. However, due to the stochastic nature of the algorithm, other movements also return probabilities (although very small). To account for this, we use a threshold value, to discard other gestures. This threshold value is obtained as the minimum probability of all the training sequences given both models (*i.e.,* the lowest recognition result for all training samples).

However, one problem arises with this configuration. Due to the state separation, *left* and *right* states are not defined in the nod HMM. Therefore, if a nodding gesture contains an unexpected leftward or rightward movement, the

recognition fails because there is no probability that the observation sequence belongs to the nodding model. A similar case happens with the recognition of the shaking gesture (*up* and *down* states are not defined in the shake HMM), the system fails to recognize a shake if the observation sequence contains an upward or downward movement.

3.3 Head Nodding and Head Shaking Recognition Using Two Complete HMMs

In a second configuration, we created two HMMs including all the possible states for nodding and shaking, like the one shown in Fig. 4. Using these HMMs, it exists a probability (although very low) that during a nodding, a *left* or *right* movement might occur. In other words, a nodding sequence can contain *left-ward* or *rightward* observations. Similarly, a shake sequence can contain *upward* or *downward* observations. With this new configuration, the recognition rate increased considerably.

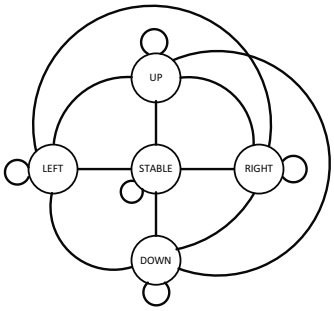

Fig. 4. Second model representation. The model is fully connected and contains all the states and observation movements.

Even though the recognition rate improved, we noticed false positives during live video testing. The false positives appeared when the person being tracked performed a simple movement such as turning right, turning left, looking up or looking down. The observation sequences from these simple movements have high probabilities given the models, *e.g.,* turning right is highly recognized as a shake; sometimes even higher than a shaking gesture. Thus, making the threshold value useless for these situations.

3.4 Head Gestures Recognition Using Six Complete HMMs

In a third configuration, we created six complete HMMs like the ones from the previous approach (Fig. 4). To train these HMMs, we collected another set of 20 videos with left and right movements. The HMMs are trained to recognize each

of the following gestures: nodding, shaking, turning left, turning right, looking up, and looking down.

With this configuration, the nodding and shaking recognition-rate improved considerably in live video, because now simple movements are recognized by their own model, minimizing the false positives. Another benefit of this added functionality is that now, we can infer additional information. For instance, we can analyze how much time the interlocutors are looking elsewhere.

4 Experimental Results

We collected a dataset of videos taken from two sources: a static camera and a wearable camera. The static camera dataset consists of 30 videos (from 10 different users) with annotated ground truth, in which half of the videos the participants are wearing sunglasses. In total, this set contains around 100 samples of each gesture: nodding, shaking, turning left, turning right, looking up, and looking down. These videos were created with two software applications: the first application displays 20 *Yes/No* questions. The user answers each question clicking a *Yes* button or a *No* button, and then the user must perform the movement (a nodding for *Yes* or a shaking for *No*). The application records the video and tracks the face for 20 frames (approx. one second) saving in a file the changes in yaw and pitch. The answers given by the user (by clicking the button) serve as ground truth for training and test. The second application requests the user to perform simple movements (turning left, turning right, looking up, and looking down), this application also records the video and tracks the head saving the yaw and pitch changes along with the ground truth (the requested movement). The wearable camera dataset consists of 10 recordings of approximately 3 minutes each, with annotated ground truth (6 gestures). These videos contain moderate ego-motion which affect the tracking.

A random 70% of the data were selected for training and the rest for test. Fig. 5 shows the head nod and shake recognition results for all the 10 videos.

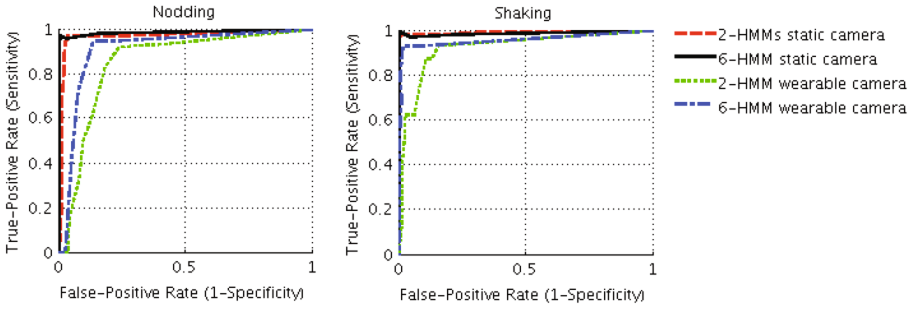

Fig. 5. Recognition curves for the static camera and the wearable camera testing data. (a) shows the head nod recognition curves when varying the discriminator threshold. (b) shows the head shake recognition curves when varying the discriminator threshold.

The ROC curves display the recognition performance of the last two HMMs configurations when varying the discriminator threshold.

Table 1 shows the area under each curve representing the recognition rate for each configuration. We can see from this table that the 6-HMMs configuration performs slightly better on the static camera videos, but performs much better on the wearable camera videos. A video of the system is available from `http://www.youtube.com/watch?v=ThvG2VOyJtE&feature=youtu.be`.

Table 1. Recognition rates for the two HMMs configurations on the static-camera test data and the wearable-camera test data

	Static camera		Wearable camera	
Gesture	2 HMMs	6 HMMs	2 HMMs	6 HMMs
Nodding	96.6%	98.5%	84.8%	90.6%
Shaking	98.7%	98.7%	92%	95.7%
Left	–	98.5%	–	95.6%
Right	–	98.3%	–	95.6%
Up	–	97.5%	–	90.3%
Down	–	97.4%	–	90.2%

5 Conclusion and Future Work

This paper presents a method for robustly recognize six head gestures (nodding, shaking, turning right, turning left, looking up, and looking down) using what we call *complete Hidden Markov Models*.

The main difference with other methods is that our HMMs consider all possible states in any given order. The selection of this approach provides great flexibility because a head nod that normally contains up and down movements, now, can also contain left and right movements as well. This is an advance over previous methods as it permits unconstrained movements of the head, while presenting robustness on video taken with wearable cameras (*e.g.,* glasses or neck-hung devices). In this case, our approach can deal with the noise introduced by ego-motion.

In future work, we will combine this method with other social interaction cues such as gaze direction and face expressions in a wearable device that will provide the user with cues indicating the level of attention or behavior of the interlocutors during social interactions. Such assistive technology could be used for instance by visually-impaired people in order to strengthen their presence and role during social meetings.

Acknowledgments. This research was partially funded by Fomix CONACYT-GDF under grant 189005 and SIP-IPN under grant 20140325.

References

1. Manduchi, R., Coughlan, J. (Computer) Vision without Sight. Commun. ACM 55(1), 96–104 (2012)
2. Krishna, S., Colbry, D., Black, J., Balasubramanian, V., Panchanathan, S., et al.: A Systematic Requirements Analysis and Development of an Assistive Device to Enhance the Social Interaction of People who are Blind or Visually Impaired. In: Workshop on Computer Vision Applications for the Visually Impaired (2008)
3. Knapp, M.L.: Nonverbal communication in human interaction. Cengage Learning (2012)
4. Wiener, W., Lawson, G.: Audition for the traveler who is visually impaired. Foundations of Orientation and Mobility 2, 104–169 (1997)
5. Dittmann, A.T., Llewellyn, L.G.: Relationship between vocalizations and head nods as listener responses. J. Pers. Soc. Psychol. 9(1), 79–84 (1968)
6. Choi, H., Rhee, P.: Head Gesture Recognition using HMMs. Expert Syst. Appl. 17(3), 213–221 (1999)
7. Kang, S.K., Chung, K.Y., Lee, J.H.: Development of head detection and tracking systems for visual surveillance. Pers. Ubiquit. Comput. 18(3), 515–522 (2014)
8. Kapoor, A., Picard, R.: A Real-Time Head Nod and Shake Detector. In: Workshop on Perceptive user interfaces, pp. 1–5. ACM (2001)
9. Tan, W., Rong, G.: A Real-Time Head Nod and Shake Detector using HMMs. Expert. Syst. Appl. 25(3), 461–466 (2003)
10. Viola, P., Jones, M.: Robust real-time object detection. Int. J. Comput. Vision 4 (2001)
11. Gunes, H., Pantic, M.: Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Allbeck, J.M., Badler, N.I., Bickmore, T.W., Pelachaud, C., Safonova, A. (eds.) IVA 2010. LNCS, vol. 6356, pp. 371–377. Springer, Heidelberg (2010)
12. Fujie, S., Ejiri, Y., Matsusaka, Y., Kikuchi, H., Kobayashi, T.: Recognition of paralinguistic information and its application to spoken dialogue system. In: Workshop on Automatic Speech Recognition and Understanding, pp. 231–236. IEEE (2003)
13. Wei, H., Scanlon, P., Li, Y., Monaghan, D.S., O'Connor, N.E.: Real-time head nod and shake detection for continuous human affect recognition. In: 14th International Workshop on Image Analysis for Multimedia Interactive Services, pp. 1–4. IEEE (2013)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
15. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. Int. J. Comput. Vision 15(1-2), 123–141 (1995)
16. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. IEEE Trans. Pattern Anal. Machine Intell. 31(4), 607–626 (2009)
17. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
18. Martins, P., Batista, J.: Monocular head pose estimation. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2008. LNCS, vol. 5112, pp. 357–368. Springer, Heidelberg (2008)
19. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)