SIViP (2011) 5:495–506
DOI 10.1007/s11760-011-0247-3

ORIGINAL PAPER

On the use of dynamic features in face biometrics: recent advances and challenges

Abdenour Hadid · Jean-Luc Dugelay · Matti Pietikäinen

Received: 1 August 2010 / Revised: 14 December 2010 / Accepted: 17 February 2011 / Published online: 29 July 2011

Abstract The way a person is moving his/her head and facial parts (such as the movements of the mouth when a person is talking) defines so called *facial dynamics* and characterizes personal behaviors. An emerging direction in automatic face analysis consists of also using such dynamic cues, in addition to facial structure, in order to enhance the performance of static image-based methods. This is inspired by psychophysical and neural studies indicating that behavioral characteristics do also provide valuable information to face analysis in the human visual system. This survey article presents the motivations, reviews the recent developments and discusses several other important issues related to the use of facial dynamics in computer vision. As a case study of using facial dynamics, two LBP-based baseline methods are considered and experimental results in different face-related problems, including face recognition, gender recognition, age estimation and ethnicity classification are reported and discussed. Furthermore, remaining challenges are highlighted and some promising directions are pointed out.

Keywords Facial dynamics · Biometrics · Spatio-temporal analysis · Behavioral features · Local binary patterns

1 Introduction

Though there has been a great deal of progress in face analysis in the last decade, many problems remain unsolved especially when dealing with illumination changes, aging, pose variations, low-image quality and occlusion [31,57]. To enhance the performance of the existing systems, a recent direction in automatic face analysis consists of using facial dynamics (the way a person is talking and moving his/her facial features, see an example of mouth movements in Fig. 1) in addition to facial structure. This emerging direction is motivated by psychophysical and neural studies (e.g., [4,19,24,46]) which indicate that, in addition to facial structure, behavioral characteristics do also provide valuable information to face analysis in the human visual system (HVS). This article looks then at the developments related to the use of facial dynamics by discussing and reviewing existing works. As a case study of using facial dynamics, baseline methods using local binary patterns (LBP) [45] are considered and experimental results in different face-related problems, including face recognition, gender recognition, age estimation and ethnicity classification are reported and discussed. Furthermore, remaining challenges are highlighted and some promising directions are pointed out.

Basically, face analysis from videos can be approached using two different strategies, depending on whether the temporal information is used or not. The most straightforward strategy applies still image-based techniques to some selected (or all) frames and then fuses the results over the sequence using for example majority voting or the weighted summation rule. Obviously, this kind of methods only exploits the abundance of frames and thus ignores the temporal information (i.e., the correlation between the frames). In contrast, the emerging strategy consists of encoding both structural and temporal information for combining facial appearance and motion. This is probably a more efficient but also a more challenging approach. Our present article focuses then on this latter category of methods which aim to exploit the temporal information in the face video sequences.

A. Hadid (✉) · M. Pietikäinen
Machine Vision Group, University of Oulu, Oulu, Finland
e-mail: hadid@ee.oulu.fi

J.-L. Dugelay
EURECOM, Sophia Antipolis, France

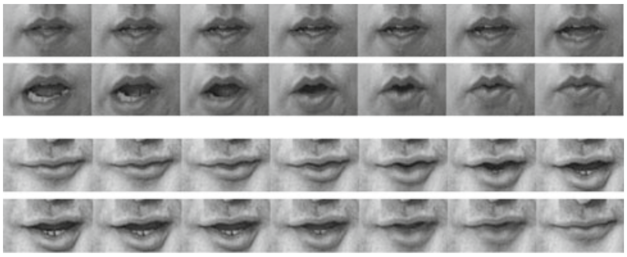

Fig. 1 Frames from two video sequences of two different persons uttering the same world "Hello", thus showing the movements around of the lip regions. In addition of being speech dependent, the lip movements are also person dependent as they convey personal characteristics

It appears that all earlier surveys on face analysis were mainly focused on still images or image sequences (e.g., [6,5,10,20,37,51,55,57]). However, during the recent years, the use of facial dynamics has gained an increased interest and this can be attested by the appearance of several special issues in journals (e.g., seeing faces in video by computers, Image and Vision Computing, 2006) and the organization of many competitions in major international conferences (such as the competition on face recognition from stills and video at ICB 2009) devoted to face analysis from videos. To this adds many other evaluations e.g., aiming at retrieving particular faces in videos (such as in TREC Video Retrieval Evaluation). All these developments have motivated us to provide the reader with this first review particularly devoted to dynamic face analysis, thus complementing the existing surveys. The aim of this article is not only to discuss the recent advances in dynamic face analysis but also to help unifying the efforts toward the development of adequate tools, protocols and databases for evaluating and monitoring the progress in dynamic face analysis. This article also aims at opening a debate on new opportunities and new challenges in the area.

The rest of this paper is organized as follows. Section 2 summarizes the main findings in psychophysics and neuroscience related to the importance of facial dynamics in the human visual system and which have direct relevance to research on automatic face analysis. Section 3 reviews some recent work attempting to combine facial appearance and dynamics for face analysis from videos. In Sect. 4, just of illustration, we describe two baseline approaches to analyze the combination of facial structure and dynamics for face analysis from videos. The first approach is using only static images and thus ignoring the facial dynamics while the second approach uses spatiotemporal representation thus combining facial structure and dynamics. We summarize the obtained experimental results in various face-related tasks. Then, we highlight some remaining challenges and point out promising directions in Sect. 5. Finally, a conclusion is drawn in Sect. 6.

2 Psychophysics of dynamic face perception

Psychological and neural studies [4,19,24,46] indicate that when people talk their changing facial expressions and head movements provide a dynamic cue for face and gender analysis. Therefore, both fixed facial features and dynamic personal characteristics are used in the human visual system to recognize and analyze faces. Among the main findings related to the importance of facial dynamics in the human visual system and which have direct relevance to research on automatic face analysis are: (a) both static and dynamic facial information are useful for face recognition and analysis; (b) people rely primarily on static information because facial dynamics provide less accurate identification information than static facial structure; (c) dynamic information contributes more to recognition under a variety of degraded viewing conditions (such as poor illumination, low-image resolution, recognition from distance etc.); (d) facial motion is learned more slowly than static facial structure; (e) recognition of familiar faces is better when they are shown as an animated sequence than as a set of multiple frames without animation. However, for unfamiliar faces, the moving sequence does not provide more useful information than multiple static images; (f) facial movement (i.e., dynamics) helps the discrimination between men and women; and (g) facial movement is fundamental to the recognition of facial expressions as analyzing an animated sequence produces more accurate results than what a collection of static images may result.

How can we interpret and exploit these findings to enhance the performance of automatic face analysis systems? A possible indication from the statements in (a), (c), (f) and (g) is that motion is a useful cue to enhance the performance of static image-based systems. Importantly, the usefulness of the motion cue increases as the viewing conditions deteriorate (statement (c)). Such an environment is often encountered in video surveillance and access control applications. Thus, an automatic recognition system should exploit both dynamic and static information. From the evidence in (d), one can interpret that facial motion is more challenging to learn and use than the face structure. Thus, a laborious training might be necessary as we move from physiological traits to behavioral ones.

3 Use of facial dynamics in face biometrics

Despite the evidences from psychophysics and neuroscience which indicate that facial movements can provide valuable information to gender classification and face recognition, only recently have researchers started to pay an important attention to the use of the facial dynamics in automatic face analysis (e.g. [29,32,34,58,60]). Unsurprisingly, the facial expression recognition problem has attained the most attention and efforts in combining facial structure and motion. This is due to the fact that facial expressions (happiness, sadness, fear, disgust, surprise and anger) are generated by contractions of facial muscles which result in temporally deformed facial features such as eye lids, lips and skin texture. Hidden Markov Models and optical flow algorithms are commonly used to determine the facial expression by modeling the dynamics of facial actions caused by skin and facial feature deformation. Complete surveys on the large number of works on facial expression recognition can be found in [10,47]. Since the role of facial dynamics in facial expression recognition is quite obvious and well studied, we focus in this article on studying the use of facial dynamics in less obvious problems such as face recognition, gender classification and age categorization.

In gender recognition, among the most notable results to date are those obtained by Moghaddam and Yang [42], and also by Baluja and Rowley [3]. Moghaddam and Yang used raw pixels as inputs to support vector machines and achieved a classification rate of 96.6% on FERET database of images scaled to 12 × 21 pixels [42]. Comparable accuracy but at a higher speed was also reported by Baluja and Rowley who used AdaBoost to combine weak classifiers, constructed using simple pixel comparisons, into a single strong classifier [3]. Note that both approaches are based on static images and assume well-aligned faces. However, in many real applications input data generally consists of video sequences and it is not always obvious to hold the face alignment assumption. One way to enhance then the performance of gender classification techniques in such environments is to design multi-modal systems combining different cues such as face, gait, facial dynamics and voice. For instance, Shan et al. investigated the fusion of face and gait at feature level and obtained performance increase when combining the two cues [53]. Naturally, in some applications such as in human–machine interaction, the gait information may not be available. While some researchers have also investigated the combination of face and voice, surprisingly very few works have addressed the combination of face appearance and facial dynamics to gender classification, despite the psychophysical studies which state that facial motion can help the discrimination between men and women [19]. Very recently, Hadid and Pietikäinen proposed a spatiotemporal approach to combine facial appearance and dynamics to gender recognition from videos, yielding interesting preliminary results [16]. The experiments showed that the combination of motion and appearance was only useful for gender analysis of familiar faces while, for unfamiliar faces, motion seemed to not provide discriminative information. In [38], Matta et al. have also explored the use of head and mouth motions in combination with facial appearance for gender recognition. Experiments on a relatively small database containing 208 video sequences of 13 different persons, showed some performance enhancement when integrating the motion information compared to the use of only facial appearance.

In contrast to other facial analysis tasks, automatic age range classification (called also age estimation or age classification) has rarely been explored, despite its vast potential applications. For instance, automatic age estimator can be very useful in smart environments where the system should adapt to the users whose behaviors and preferences are different at different ages. Among the few works on age estimation, yet based only on static images, are those of Lanitis et al. [26] and Geng et al. [13]. Lanitis et al. used a simple quadratic aging function to model the relation between face and age, while Geng et al. modeled the sequence of a particular individual's face images sorted in time order by a subspace in which unseen faces are then projected for age estimation. Unfortunately, due to its challenging nature and lack of clear psychophysical evidences, no work has yet clearly addressed the use of facial dynamics in age classification.

In face recognition, there have been many attempts to exploit the facial dynamics. Perhaps, the most popular approach to model temporal and spatial information is based on the hidden Markov models (HMM) which have been applied to face recognition from videos e.g., in [34]. The principle of using HMMs for dynamic face recognition is quite simple: during the training phase, an HMM is created to learn both the statistics and temporal dynamics of each individual. During the recognition process, the temporal characteristic of the face sequence is analyzed over time by the HMM corresponding to each subject. The likelihood scores provided by the HMMs are compared. The highest score provides the identity of a face in the video sequence.

In the following, we review in more details the main attempts to explore the use of facial dynamics for face recognition from videos.

3.1 Hidden Markov models

Hidden Markov models are powerful tools to model temporal motion information. For instance, they have been successfully used in speech recognition, gesture and facial expression recognition. In [21], Huang and Trivedi were among the first to adopt them for modeling the temporal motion information in the video sequences for dynamic face recognition. However, the authors employed HMMs with a single state and a single Gaussian component which is equivalent to using a single multidimensional Gaussian approximation. In such a configuration, the temporal correlation is unfortunately not adequately exploited.

Later, Liu and Cheng [34] extended the work of Huang and Trivedi by successfully applying HMMs for temporal face video recognition. To avoid singularities on the estimation of the covariance matrices, Liu and Cheng modified the training algorithm as follows: each covariance matrix was gradually adapted from a global diagonal one (a general model) by using its class-dependent data. Liu and Cheng also proposed an online version of their recognition system, by implementing an adaptive strategy for the HMMs. More precisely, each test sequence successfully recognized was used to update the model parameters of the client in question, by applying a maximum a posteriori (MAP) adaptation technique. In order to discard miss-classified testing videos, the likelihood difference values were used as confidence measures. In conclusion, the system exploiting adaptive HMMs performed better than the one without adaptation, and both obtained higher recognition scores than the eigenface approach with majority voting.

Recently, Tistarelli et al. have also adopted an HMM-based approach to capture both the face appearance and the face dynamics for face recognition from videos [54]. The authors proposed a multidimensional extension of HMMs, called Pseudo Hierarchical HMM, in which the emission probability of each state is represented by another HMM, while the number of states is determined from the data by unsupervised clustering of facial motion in the video. The method was tested on a limited homemade database of 21 subjects, showing encouraging performance compared to PCA and other static image-based methods using HMMs.

3.2 Discriminant analysis on facial optical flow

In [7], Chen et al. explored dynamic face recognition by considering the optical flow of the facial movements as a biometric cue for person recognition from video. The algorithm applies the Fisherface method on facial motion. For each video, it firstly calculates a sequence of optical flow fields and subsequently concatenates them (frame by frame) to form a unique high-dimensional vector. Then, the motion vectors are projected into a discriminative feature subspace, obtained by applying PCA and LDA on the training data set. Finally, the system recognizes identities using a nearest neighbor classifier. Under illumination changes, the experiments showed better recognition performance compared to the original Fisherface approach. However, temporal segmentation and video chunk normalization were not explicitly addressed by the authors, assuming having sequences of commensurate facial motion. Every frame was also semiautomatically preprocessed before the optical flow computation, in order to align and normalize the head size and location. Furthermore, only the lowest half of optical flow fields was used to extract features for recognition, so that these were mostly related to mouth motion.

3.3 Stochastic tracking and recognition through particle filtering

Stochastic tracking and recognition approaches are based on a unified probabilistic framework, in which individuals are simultaneously tracked and recognized by estimating the posterior probability density function of a time series state space model (TSSSM). Tracking is formulated as a Bayesian inference problem, and it is solved as a probability density propagation problem (due to the temporal nature of tracking itself). Recognition is obtained by applying the MAP rule on the posterior probabilities. A TSSSM with nonlinear dynamics and non-Gaussian noise model is generally adopted. Its state and probability estimations are numerically computed using sequential Monte Carlo methods [9,33], and more particularly the sequential importance sampling (SIS) algorithm.

In [30], Li and Chellappa were the first to develop a generic approach for stochastic tracking and verification using particle filtering. They implemented a simplified TSSSM with no identity variable, in which only the tracking motion vector was estimated and propagated. They also proposed two facial representations for the observations: the common intensity images of the face, and an EGM (elastic graph matching) based representation of the facial landmarks.

Then, Zhou et al. [61] improved the approach of Li and Chellappa by including both the tracking motion vector and the identity variable in the TSSSM. They also introduced a new observation likelihood by explicitly modeling the appearance changes within videos using a truncated Laplacian and the intrapersonal appearance variations using a probabilistic subspace density, proposed by Moghaddam [41]. More interestingly, the authors have also developed a probabilistic learning approach to automatically build user models from video frames. During the enrollment phase, the algorithm incrementally selected exemplar frames of an individual and used them as mixture centers of a probabilistic distribution for that client. In the recognition phase, they modified the TSSSM and the observation likelihood accordingly, by adding the exemplar variable in the state space model. The authors obtained very good identification rates on the small (29 subjects) motion of body video database [14].

Later, Zhou et al. [59] refined their previous approach by deriving an adaptive version. They modified the observation likelihood by modeling the appearance changes within videos using an adaptive appearance model, the intra and inter personal appearance variations using a probabilistic subspace density [41], and up weighting frontal view frames using another probabilistic subspace density. Then, the authors proposed an adaptive motion model, which consisted of an adaptive velocity model (predicted using a first-order linear approximation), an adaptive noise component (function of the prediction error) and an adaptive number of particles (in the SIS algorithm). Moreover, they included an occlusion handling technique based on robust statistics, which stopped the automatic adaptations during occluded frames. The approach yielded in best performance compared to all earlier stochastic methods on the motion of body video database [14].

3.4 Tracking and recognition using probabilistic appearance manifolds

Tracking and recognition have often been considered as two independent components of video-based person recognition systems. However, many recent works have proposed to integrate these two tasks into a single framework as shown above with the use of TSSSM. An alternative to simultaneously track and recognize individuals is the probabilistic appearance manifold approach [28] which is an extension to video tracking and recognition of the concept of appearance manifold, introduced by Murase and Nayar [43]. For instance, in [28], Lee et al. developed a probabilistic appearance manifold approach for person tracking and recognition from video sequences. The authors applied Bayesian inference to include the temporal coherence of human motion in the distance calculation. They replaced the conditional probability by using the joint conditional probabilities, which were recursively estimated using the transitions between sub-manifolds. In experiments using a small database of 20 individuals, the approach outperformed many conventional image-based recognition techniques and other approaches without temporal coherence. The system was also able to detect identity changes and to handle large pose variations.

3.5 Gaussian mixture modeling on unconstrained head motion

In [35], Matta and Dugelay explored the use of head and facial motion for person recognition from video sequences. They presented a person recognition system that exploited the unconstrained head motion information extracted by tracking a few facial landmarks in the image plane. In particular, each video sequence was firstly preprocessed by semiautomatically detecting the face which was then automatically tracked by following a few facial landmarks over time using a template matching strategy. Then, the extracted patterns were geometrically normalized in order to calculate discriminative feature vectors, which were successively used to estimate the client models through a Gaussian mixture model (GMM) approximation. Person identification and verification were finally performed by applying the probability theory and the Bayesian decision rule (also called Bayesian inference).

Afterwards, Matta et al. proposed a multimodal extension of their person recognition system [36,50]. They successfully integrated the head motion information with mouth motion and facial appearance, by taking advantage of a unified probabilistic framework. The authors developed a new temporal subsystem that had an extended feature space enriched by additional mouth parameters [50]. They also introduced a complementary spatial subsystem based on a probabilistic extension of the original eigenface approach [36]. In the end, Matta et al. derived an integration scheme to combine the similarity scores of the two parallel subsystems using optimal score fusion strategy [36]. The authors reported very good results on a limited homemade database called The Italian TV speakers video database (see Sect. 5.1 for database description).

3.6 Volumetric features for combining motion and appearance

Recent developments also showed that volumetric features can be use to encode both facial motion and appearance [17,56]. The idea is to consider a face video sequence as a rectangular prism from which volumetric primitives can be collected into a histogram representing the appearance and motion of the face in the video sequence. Following these lines, Hadid and Pietikäinen explored a volume local LBP-based spatiotemporal representation for face recognition from videos with very good results [17]. Starting from the observation that the volumetric features consist of both intra and extra-personal information (corresponding to both facial expression and identity), they proposed a robust recognition system using AdaBoost learning. The idea was to classify the volumetric facial patterns into intra and extra classes, and then use only the extra-class features for recognition. Experiments on MoBo video face database [14] showed significant increase in the recognition rates compared to many static image-based methods. The principle of using volumetric features for combining motion and appearance for dynamic face recognition is further detailed in Sect. 4.2.

4 Case study: experimental analysis on encoding facial dynamics using local binary patterns and SVMs

To gain insight into the use of facial dynamics, two baseline approaches based on LBP features [45] and support vector machines (SVM) are implemented and discussed in this section. The first approach is using only static images and thus ignoring the facial dynamics while the second approach uses spatiotemporal representation thus combining facial structure and dynamics. The aim of the experiments is to evaluate the benefit of incorporating the facial dynamics. The choice of adopting LBP approach [45] is motivated by the recent success of using it for combining appearance and motion for face and facial expression recognition [17,56] and also for dynamic texture recognition [56]. We describe below the two experimental approaches and then report the experimental results on face recognition, gender classification, age estimation and ethnicity classification.

4.1 Static image-based approach using LBP

The LBP texture analysis operator, introduced by Ojala et al. [44,45], is defined as a grayscale invariant texture measure, derived from a general definition of texture in a local neighborhood. It is a powerful means of texture description and among its properties in real-world applications are its discriminative power, computational simplicity and tolerance against monotonic grayscale changes caused, e.g., by illumination variations. LBP can be efficiently used for representing and analyzing faces in both still images and video sequences [2,17,56].

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas etc. The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \tag{1}$$

where g_c corresponds to the gray value of the center pixel (x_c, y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities. For an efficient representation, facial images are first divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Figure 2 shows an example of an LBP-based facial representation. In such a description, the face is represented in three different levels of locality: the LBP labels for the histogram contain information about the patterns on a pixel level, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face. This locality property, in addition to the computational simplicity and tolerance against illumination changes, are behind the success of LBP approach for facial image analysis [2].

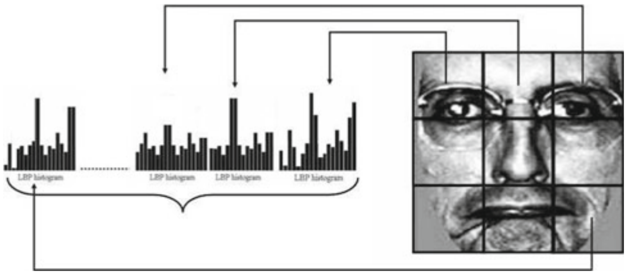

Fig. 2 Example of an LBP-based facial representation

Given a target face video sequence, a straightforward approach to perform recognition or classification is to analyze each frame and then combine the results through majority voting which consists of determining the gender (or age or identity) in every frame and then fusing the results. Therefore, each facial image (frame) is divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Then, the results are presented to an SVM classifier for recognition. Finally, the recognition scores over the face sequence are fused using majority voting. In such an approach, only static information is used while the facial dynamics are discarded.

4.2 Spatiotemporal-based approach using volume LBP

For spatiotemporal representation, volume LBP (VLBP) operator has been introduced in [56] and successfully used for combining appearance and motion for face and facial expression recognition [17,56] and also for dynamic texture recognition [56]. The idea behind VLBP is very simple. It consists of looking at a face sequence as a rectangular prism (or volume) and defining the neighborhood of each pixel in three-dimensional space (X, Y, T) where X and Y denote the spatial coordinates and T denotes the frame index (time). Then, similarly to LBP in spatial domain, volume textons can be defined and extracted into histograms. Therefore, VLBP combines structure and motion together to describe the moving faces. Figure 3 explains the principle of rectangular prism and shows an example of VLBP based representation of a face sequence.

Once the neighborhood function is defined, each face sequence can be divided into several overlapping rectangular prisms of different sizes, from which local histograms of VLBP code occurrences are extracted. Then, instead of simply concatenating the local histograms into a single histogram, AdaBoost learning algorithm [12] is adopted for automatically determining the optimal size

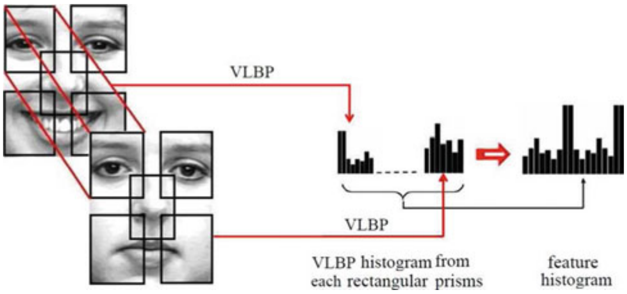

Fig. 3 Example of a VLBP-based representation of a face sequence

and locations of the local rectangular prisms, and more importantly for selecting the most discriminative VLBP patterns for classification while discarding the features which may hinder the classification process. So, to combine facial structure and dynamics, VLBP features are firstly extracted from the face sequences and feature selection is performed using AdaBoost. The result is then fed to an SVM for classification. In such an approach, both static facial information and facial dynamics are used.

4.3 Experiments on face recognition

The two experimental approaches are first applied to the problem of face recognition from videos. In order to experiment with a large amount of facial dynamics, resulted for example from the movements of the facial features when the individuals are talking, CRIM video database [8] is considered. This is large set of 591 face sequences showing 20 persons reading broadcast news for a total of about 5 h. The database is originally collected for audio–visual recognition. There are between 23 and 47 video sequences for each individual. The size of the extracted face images is 130 × 150 pixels. Half of the face sequences of each subject is randomly selected for training while the other half is used for testing. We report the average recognition rates of 100 random permutations. The performances of both static image-based and spatiotemporal-based approaches on CRIM video database are shown in Table 1. From the results, we can notice that the spatiotemporal based method (i.e., combination of face structure and dynamics) significantly outperforms the static image-based method (i.e., using only facial structure). The better performance of the spatiotemporal method is in agreement with the neuropsychological evidence [46] stating that facial dynamics are useful for face recognition.

4.4 Experiments on gender recognition

Determining whether the person whose face is in the given video is a man or a woman is useful for many applications such as more affective human–machine interaction, restricting access to certain areas based on gender, collecting demographic information in public places, counting the number of women entering a retail store and so on. Similarly to the face recognition experiments, the static image-based and the spatiotemporal-based approaches are adapted and applied to the problem of gender recognition from videos. Three different publicly available video face databases (namely CRIM [8], VidTIMIT [52] and Cohn-Kanade [22]) are considered. They contain a balanced number of male's and female's sequences and include several subjects moving their facial features by uttering phrases, reading broadcast news or expressing emotions. The datasets are randomly segmented to extract over 4,000 video shots of 15–300 frames each. From each shot or sequence, the eye positions are automatically detected from the first frame. The determined eye positions are then used to crop the facial area in the whole sequence. Finally, the resulted images are scaled into 40 × 40 pixels. For evaluation, a fivefold cross validation test scheme is adopted by dividing the 4,000 sequences into five groups and using the data from four groups for training and the left group for testing. This process is repeated five times and we report the average classification rates. When dividing the data into training and test sets, we explicitly considered two scenarios. In the first one, a same person may appear in both training and test sets with face sequences completely different in the two sets due to facial expression, lighting, facial pose etc. The goal of this scenario is to analyze the performance of the methods in determining the gender of familiar persons seen under different conditions. In the second scenario, the test set consists only of persons who are not included in the training sets. This is equivalent to train the system on one or more databases and then do evaluation on other (different) databases. The goal of this scenario is to test the generalization ability of the methods to determine the gender of unseen persons.

Table 1 Average face recognition rates using static image-based and spatiotemporal-based approaches on CRIM video database [8]

Method	Average face recognition rate
Static image-based approach	93.3%
Spatiotemporal-based approach	98.1%

Table 2 summarizes the gender classification results using the two approaches (static image based and spatiotemporal based) in both scenarios (familiar and unfamiliar). We can notice that both methods gave better results with familiar faces than unfamiliar ones. This is not surprising and can be explained by the fact that perhaps the methods did not rely only on gender features for classification but may also exploited information about face identity. For familiar faces, the combination of facial structure and dynamics yielded in

Table 2 Gender classification results on test videos of familiar and unfamiliar subjects using static image-based and spatiotemporal-based methods

Method	Gender classification rate	
	Familiar subjects	Unfamiliar subjects
Static image-based	94.4%	90.6%
Spatiotemporal-based	100%	82.9%

perfect classification rate of 100%. This proves that the system succeeded in learning and recognizing the facial behaviors of the subjects even under different conditions of facial expression, lighting and facial pose. For unfamiliar faces, the combination of facial structure and dynamics yielded in classification rate of about 83% which is still encouraging although the best result for unfamiliar faces is obtained using the static image-based approach (without facial dynamics). This may indicate that incorporating motion information with facial appearance was useful for only familiar faces but not with unfamiliar ones. More detailed experiments and results can be found in [15].

4.5 Experiments on age estimation

Automatic age estimation (or classification) aims at determining the age range of a target face. This is a very challenging problem but also a very useful application. To study whether facial dynamics may enhance the automatic age estimation performance, a set of experiments using the static image-based and spatiotemporal-based approaches is performed. Five age classes are considered as follows: child = 0–9 years old; youth = 10–19; adult = 20–39; middle age = 40–59 and elderly = above 60. Then, a novel classification scheme based on a tree of four SVM classifiers is built. The first SVM classifier is trained to learn the discrimination between `child` class and the rest. If the target face is assigned into the `child` category, then the classification is completed. Otherwise, the second SVM classifier is examined to decide whether the face belongs to the `Youth` category or not. If not, the third SVM is examined and so on.

The static image-based and spatiotemporal-based approaches are applied to age estimation from videos. For evaluation, a set of video sequences (mainly showing celebrities giving speeches in TV programs and News) is collected from Internet. The videos of unknown individuals (especially children), are manually labeled using our (human) perception of age. Then, the videos are randomly segmented to extract about 2,000 video shots of about 300 frames each. In the experiments, we adopted a tenfold cross validation test scheme by dividing the 2,000 sequences into 10 groups and using the data from 9 groups for training and the left group for testing. We repeated this process 10 times and we report the average classification rates.

Table 3 Average age classification rates

Method	Average age classification rate
Static image-based approach	77.4%
Spatiotemporal-based approach	69.2%

Table 4 Average ethnicity classification rates using static image-based and spatiotemporal-based approaches

Method	Ethnicity classification rate
Static image-based approach	97.0%
Spatiotemporal-based approach	99.2%

The performances of both static image-based and spatiotemporal-based approaches are shown in Table 3. From the results, we can notice that both methods did not perform very well and this somehow confirms the difficulty of the age estimation problem. Interestingly, the static information based method significantly outperformed the spatiotemporal based method (i.e., combination of face structure and dynamics). This might be an indication that facial dynamics is not useful for age estimation. However, due to the challenging nature of the age estimation problem, it is perhaps too early to make such a conclusion and more investigations are needed to study the integration of facial dynamics and facial structure for age estimation.

4.6 Experiments on ethnicity classification (Asian vs. non-Asian)

Similarly to the previous experiments on gender recognition, the two experimental approaches are also applied to ethnicity classification from videos. Because of lack of ground truth data for training, only two ethnic classes (namely Asian and non-Asian) are considered. The same set of 2,000 video shots previously used in the experiments on age estimation is also considered here for ethnicity classification tests. A manual labeling yielded in 81% of non-Asian and 19% of Asian data samples (i.e., video shots). For evaluation, we also adopted a fivefold cross validation test scheme.

The performances of both static image-based and spatiotemporal-based approaches are shown in Table 4. From the results, we can notice that both approaches perform quite very well but the spatiotemporal-based method (i.e., combination of face structure and dynamics) slightly outperforms the static image-based method (using only facial structure). This is somehow surprising because one may not expect better results using spatiotemporal methods for ethnicity classification.

4.7 Discussion

To gain insight into the use facial dynamics in face biometrics, we considered two approaches to face analysis from videos using LBP features and SVMs, and reported preliminary experimental results on several problems including face recognition, gender classification, age estimation and ethnicity determination. The experiments results on face recognition showed that the spatiotemporal-based method significantly outperforms the static image-based method. This is somehow in agreement with the neuropsychological evidence [46] stating that facial dynamics are useful for face recognition. The experiments on age estimation pointed out that combining face structure and dynamics does not enhance the performance of static image-based automatic systems. Our experiments also showed that incorporating motion information with facial appearance for gender classification might be only useful for familiar faces but not with unfamiliar ones (while the psychological and neural studies indicated that facial movements do contribute to gender classification in the HVS). Finally, our experiments on the ethnicity classification problem yielded in quite surprising results indicating some relative usefulness of facial dynamics in ethnicity classification.

The primary goal of the above experiments is to provide the reader with clear case studies and examples on using facial dynamics for face analysis from videos. Note that the reported results may be specific to the methods and test material that were used in the experiments. Therefore, it is perhaps too early to make final conclusions on the role of facial dynamics in video-based face analysis and face biometrics. We strongly believe that our reported experiments and results will advance and stimulate the ongoing efforts among the research community.

5 Issues, challenges and future directions

5.1 Lack of standard databases and protocols

There exist several publically available face video databases but unfortunately none of them is particularly intended for studying facial dynamics for face biometrics. The lack of standard databases and associated protocols makes the evaluation of the progress in the field of behavioral face analysis very difficult. So far, the proposed algorithms have been tested on databases which are recorded for other purposes. For instance, many works in video-based face recognition research have considered the motion of body (MoBo) video database [14] although it was originally collected for the purposes of human identification and activity recognition from distance. The MoBo database contains video sequences of 29 different subjects walking on a treadmill. Four different walking situations are considered: slow walking, fast walking, incline walking and carrying a ball. Some examples of frames are shown in Fig. 4.

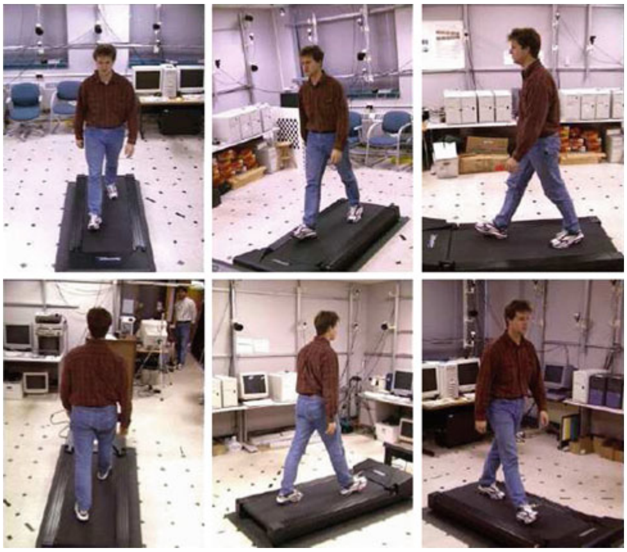

Fig. 4 Examples of frames from the MoBo database [14]

The Italian TV speakers video database is also widely used by Matta et al. in their experiments on head and facial motion analysis for face and gender recognition [35,36,50]. The database contains 208 video clips of 13 TV speakers (8 men and 5 women) from the Italian national channel RAI 1. The videos present TV speakers announcing the news of the day. A typical sequence has a spatial resolution of 352 × 288 pixels and a temporal resolution of 23.97 frames per second, and lasts 13 seconds. The videos are of low quality, acquired using a fixed camera and compressed at 118 Kbits per second, and they have been collected during a period of 21 months. Figure 5 shows some frames of four different subjects. Unfortunately, this database is not publicly available for research purposes. Another drawback of the database lies in its small size.

There are several other databases containing face video sequences that were used in different works involving facial dynamic analysis. Among these databases are Secure phone PDA [25], M2VTS [48], XM2VTS [40], BANCA [49], VidTIMIT [52], AVICAR [27], CRIM [11], BIOMET [23], M3 Corpus [39], BioSecure [1] etc. Most of these databases were originally captured for evaluating multimodal biometric systems and thus are not optimal for studying the use of facial behavior. Some of these databases are publicly available for research purposes while others are not. Even when evaluated on the same database, it is difficult to objectively compare the performance of different techniques because of lack of standards, protocols (e.g., which videos to use for training and testing?) and ground truth (e.g., the unavailability of age information for experiments on age classification).

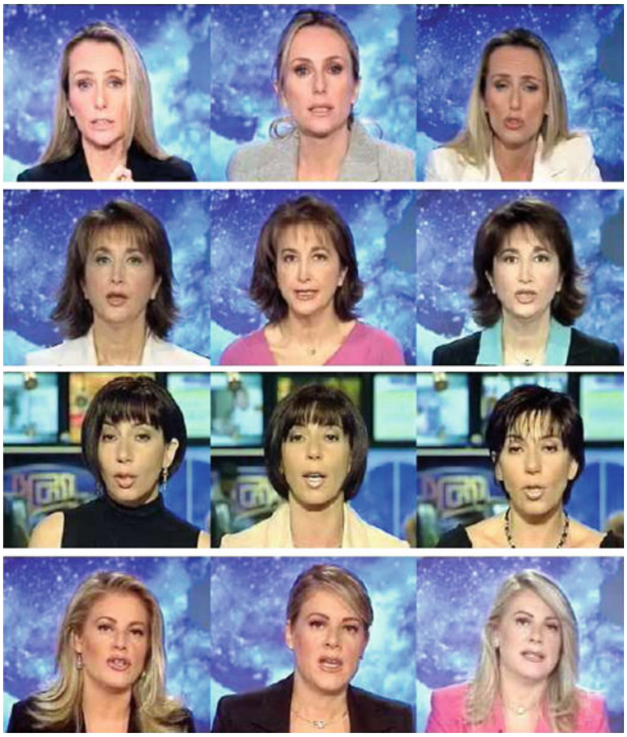

Fig. 5 Examples of variations in the Italian TV speakers video database

To advance the research efforts and track the progress in the field of facial dynamics analysis, we are planning in the near future to annotate few existing face video databases and define clear protocols that we will publically release and share within the research community. Our next goal is to capture a new database particularly dedicated to the use of facial dynamics and publically release it with associated protocols and baseline methods. We are also aiming at organizing international competitions to evaluate the progress on these databases and protocols.

5.2 Online adaptation

An automatic recognition system should exploit both facial dynamic and static information. However, these two cues do not equally contribute to recognition as the role of motion depends on a number of factors such as the familiarity of the faces, the amount of motion, the viewing conditions, etc. Thus, depending on the situation, the automatic system should bias the role of each cue rather than integrate them in an ad hoc manner (i.e., with fixed weights). For instance, the system should increase the contribution of the facial dynamics for low-resolution images and decrease this contribution for higher image resolution (like it is done in the human visual system). However, in most existing techniques, the dynamic cue is not automatically adapted (not biased) to the given situation. Continuous adaption of the contribution of the facial dynamics and structure remains an open issue that should be carefully addressed in the future efforts to advance video based face analysis research. This also suggests that the existing works have not yet shown their full potential and need further investigation.

5.3 Feature selection

Facial dynamics convey information concerning not only the identity of the subject but also the facial expression, the emotion, the gender etc. Therefore, not all the facial dynamics are useful for face recognition, for instance. This means that some part of the dynamic information is useful for recognition while another part may also hinder the recognition. Obviously, the useful part is that defining the extra-personal characteristics while the non-useful part concerns the intra-class information such as facial expressions and emotions. For recognition, one should then select only the extra-personal characteristics. Selecting the useful facial dynamics is an open issue that should be carefully investigated. It has been partially addressed by Hadid and Pietikäinen in [16].

5.4 Automatic tracking, normalization and recognition

Video-based systems for face analysis should perform automatic tracking, feature extraction and normalization and then recognition. However, most methods can only handle well-aligned faces while many others are based on the manual or semiautomatic normalization of the features before recognition. This obviously limits their use in practical scenarios and uncontrolled conditions. The recent works on integrating tracking and recognition into a single framework (e.g., with the use of TSSSM) have shown to provide interesting results and this should be further extended to develop fully automatic systems.

6 Conclusion

In this work, we discussed the psychological and neural findings about the importance of facial dynamics in the human visual system and reviewed the major attempts to combine facial structure and motion for automatic face analysis. Because finding efficient representations for combining facial structure and dynamics for face analysis from videos is challenging, most of the existing works limit the scope of the problem by discarding the facial dynamics and only considering the structure. To gain insight into the use of facial dynamics in face biometrics, we provided the reader with clear case studies and experimental methods and results on several problems including face recognition, gender classification, age estimation and ethnicity determination from video sequences.

Unfortunately, most of the methods which use spatiotemporal representations for face recognition from videos have not yet shown their full potential as they suffer from different drawbacks such as the use of only global features while local information is shown to also be important to facial image analysis [18] and the lack of discriminating between the facial dynamics which are useful for recognition from those which can hinder the recognition process. In addition, most methods can only handle well-aligned faces thus limiting their use in practical scenarios and uncontrolled conditions.

We also pointed out the urgent need for unifying the efforts to develop standard databases and associated protocols for evaluating and monitoring the progress in dynamic face analysis. Among the highlighted challenges are the online adaption of the contribution of the facial dynamics and structure which remains an open issue, the need of performing feature selection for using only the useful facial dynamics, and the importance of developing fully automatic systems exploiting facial dynamics for practical scenarios.

Our results suggest that existing representations for combining facial structure and dynamics have not yet shown their full potential and need further investigation. By considering the human visual system as a valuable source of inspiration, the aim of dynamic face analysis is to enhance the performance of the existing systems and develop new technologies which can efficiently detect, track, recognize, analyze faces from videos and, why not, also mimic most of the remarkable abilities of the human visual system. Although we are not there yet, current progress in the field gives us much reason to be optimistic.

Acknowledgments The financial support of the Academy of Finland is gratefully acknowledged. This work has been partially performed within the TABULA RASA project 7th Framework Research Program of the European Union (EU), grant agreement number: 257289.

References

1. URL http://www.biosecure.info/
2. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
3. Baluja, S., Rowley, H.: Boosting sex identification performance. Int. J. Comput. Vis. **71**, 111–119 (2007)
4. Bassili, J.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. J. Pers. Soc. Psychol. **37**, 2049–2059 (1979)
5. Bowyer, K., Chang, K., Flynn, P.: A survey of approaches to three-dimensional face recognition. In: International Conference on Automatic Face and Gesture Recognition, vol. 1, pp. 358–361 (2004)
6. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: a survey. In: IEEE, vol. 83, no. 5, pp. 705–740 (1995)
7. Chen, L.F., Liao, H.-Y.M., Lin, J.C.: Person identification using facial motion. In: International Conference on Image Processing, 2001, vol. 2, pp. 677–680 (2001)
8. CRIM: http://www.crim.ca/
9. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. Stat. Comput. **10**(3), 197–208 (2000)
10. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**, 259–275 (2003)
11. Foucher, S., Lalibert, T.F., Boulianne, G., Gagnon, L.: A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP 2006), pp. I597–I600 (2006)
12. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci **55**(1), 119–139 (1997)
13. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2234–2240 (2007)
14. Gross, R., Shi, J.: The CMU motion of body (mobo) database. Technical report (2001)
15. Hadid, A., Pietikäinen, M.: Combining motion and appearance for gender classification from video sequences. In: 19th International Conference on Pattern Recognition (ICPR 2008), p. 4 (2008)
16. Hadid, A., Pietikäinen, M.: Combining appearance and motion for face and gender recognition from videos. Pattern Recognit. **42**(11), 2818–2827 (2009)
17. Hadid, A., Pietikäinen, M., Li, S.Z.: Learning personal specific facial dynamics for face recognition from videos. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (in conjunction with ICCV 2007), pp. 1–15 (2007)
18. Heisele, B., Ho, P., Wu, J., Poggio, T.: Face recognition: component based versus global approaches. Comput. Vis. Image Underst. **91**(1–2), 6–21 (2003)
19. Hill, H., Johnston, A.: Categorizing sex and identity from the biological motion of faces. Curr. Biol. **11**(11), 880–885 (2001)
20. Hjelmas, E., Low, B.K.: Face detection: a survey. Comput. Vis. Image Underst. **83**(3), 236–274 (2001)
21. Huang, K.S., Trivedi, M.M.: Streaming face recognition using multicamera video arrays. International Conference on Pattern Recognition, vol. 4, p. 40213 (2002)
22. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
23. Kittler, J., Nixon, M.S. (eds.): Audio- and video-based biometrie person authentication. Lecture Notes in Computer Science, vol. 2688. Springer (2003)
24. Knight, B., Johnston, A.: The role of movement in face recognition. Vis. Cognit. **4**, 265–274 (1997)
25. Koreman, J., Morris, A., Jassim, S., Sellahewa, H., Chollet, G., Aversano, G., Salicetti, S., Allano, L. (eds.): Multi-modal biometric authentication on the Secure Phone PDA (2006)
26. Lanitis, A., Taylor, C., Cootes, T.: Towards automatic simulation of aging affects on face images. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 442–455 (2002)
27. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T.: Avicar: Audio-visual speech corpus in a car environment. In: Proceedings of International Conference on Spoken Language (2004)
28. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Visual tracking and recognition using probabilistic appearance manifolds. Comput. Vis. Image Underst. **99**(3), 303–331 (2005)
29. Li, B., Chellappa, R.: Face verification through tracking facial features. J. Opt. Soc. Am. **18**, 2969–2981 (2001)

30. Li, B., Chellappa, R.: A generic approach to simultaneous tracking and verification in video. IEEE Trans. Image Process. **11**(5), 530–544 (2002)
31. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005)
32. Li, Y.: Dynamic face models: construction and applications. Ph.D. thesis, Queen Mary, University of London (2001)
33. Liu, J.S., Chen, R.: Sequential monte carlo methods for dynamic systems. J. Am. Stat. Assoc. **93**, 1032–1044 (1998)
34. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 340–345 (2003)
35. Matta, F., Dugelay, J.L.: Person recognition using human head motion information. In: Articulated Motion and Deformable Objects, pp. 326–335 (2006)
36. Matta, F., Dugelay, J.L.: Video face recognition: a physiological and behavioural multimodal approach. In: ICIP 2007, 14th IEEE International Conference on Image Processing, 16–19 Sept 2007, San Antonio, USA (2007)
37. Matta, F., Dugelay, J.L.: Person recognition using facial video information: a state of the art. Image Vis. Comput. **20**(3), 180–187 (2009)
38. Matta, F., Saeed, U., Mallauran, C., Dugelay, J.L.: Facial gender recognition using multiple sources of visual information. In: MMSP 2008, 10th IEEE International Workshop on MultiMedia Signal Processing, 8–10 Oct 2008, Cairns, Queensland, Australia (2008)
39. Meng, H., Ching, P.C., Lee, T., Mak, M.W., Mak, B., Moon, Y.S., Siu, M.H., Tang, X., Hui, H., Lee, A., Lo, W.K., Ma, B. (eds.): The Multi-Biometric, Multi-Device and MultiLingual (M3) Corpus (2006)
40. Messer, K., Matas, J., Kittler, J., Lüttin, J., Maitre, G.: XM2VTSDB: The extended M2VTS database. In: Audio- and Video-Based Biometric Person Authentication, AVBPA'99, March 1999, 16 IDIAP–RR 99-02, pp. 72–77. Washington, DC (1999)
41. Moghaddam, B.: Principal manifolds and probabilistic subspaces for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 780–788 (2002)
42. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 707–711 (2002)
43. Murase, H., Nayar, S.K.: Visual learning and recognition of 3-d objects from appearance. Int. J. Comput. Vis. **14**(1), 5–24 (1995)
44. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognit. **29**, 51–59 (1996)
45. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 971–987 (2002)
46. O'Toole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: a psychological and neural synthesis. Trends Cognit. Sci. **6**, 261–266 (2002)
47. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12), 1424–1455 (2000)
48. Pigeon, S., Vandendorpe, L.: The M2VTS multimodal face database (release 1.00). In: AVBPA'97: Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 403–409. Springer, London, UK (1997)
49. Popovici, V., Thiran, J., Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Ruiz, B., Poiree, F.: The BANCA database and evaluation protocol. In: 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK. Lecture Notes in Computer Science, vol. 2688, pp. 625–638. SPIE (2003)
50. Saeed, U., Matta, F., Dugelay, J.L.: Person recognition based on head and mouth dynamics. In: MMSP 2006, IEEE International Workshop on Multimedia Signal Processing, 3–6 Oct 2006, Victoria, Canada (2006)
51. Samal, A., Iyengar, P.: Automatic recognition and analysis of human faces and facial expression: a survey. Pattern Recognit. **25**(1), 65–77 (1992)
52. Sanderson, C., Paliwal, K.K.: Noise compensation in a person verification system using face and multiple speech feature. Pattern Recognit. **36**(2), 293–302 (2003)
53. Shan, C., Gong, S., McOwan, P.: Learning gender from human gaits and faces. In: IEEE International Conference on Advanced Video and Signal based Surveillance, pp. 505–510 (2007)
54. Tistarelli, M., Bicego, M., Grosso, E.: Dynamic face recognition: From human to machine vision. Image Vis. Comput. **27**(3), 222–232 (2009)
55. Yang, M.H., Kriegman, D.J., Ahuja, N.: Detecting faces in images: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **24**, 34–58 (2002)
56. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
57. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **34**(4), 399–458 (2003)
58. Zhou, S., Chellappa, R.: Probabilistic human recognition from video. In: European Conference on Computer Vision, pp. 681–697 (2002)
59. Zhou, S., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Trans. Image Process. **13**, 1434–1456 (2004)
60. Zhou, S., Krueger, V., Chellappa, R.: Face recognition from video: a condensation approach. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 221–228 (2002)
61. Zhou, S., Kruger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Comput. Vis. Image Underst. **91**, 214–245 (2002)