Monocular 3D Facial Information Retrieval for Automated Facial Expression Analysis

Meshia Cédric Oveneke *, Isabel Gonzalez *, Weiyi Wang *, Dongmei Jiang †, and Hichem Sahli *‡

* VUB-NPU Joint AVSP Research Lab, Vrije Universiteit Brussel (VUB), Department of Electronics & Informatics (ETRO), Pleinlaan 2, 1050 Brussel, Belgium. Email: {mcovenek,igonzalez,wwang}@etro.vub.ac.be

† VUB-NPU Joint AVSP Research Lab, Northwestern Polytechnical University (NPU), Shaanxi Key Lab on Speech and Image Information Processing, 127 Youyi Xilu, X’ian 710072, China. Email: jiangdgm@nwpu.edu.cn

‡ Interuniversity Microelectronics Centre (IMEC) Kepeldreef 75, 3001 Heverlee, Belgium. Email: hsahli@vub.ac.be

Abstract—Understanding social signals is a very important aspect of human communication and interaction and has therefore attracted increased attention from various research areas. Among the different types of social signals, particular attention has been paid to facial expression of emotions and its automated analysis from image sequences. Automated facial expression analysis is a very challenging task due to the complex three-dimensional deformation and motion of the face associated to the facial expressions and the loss of 3D information during the image formation process. As a consequence, retrieving 3D spatio-temporal facial information from image sequences is essential for automated facial expression analysis. In this paper, we propose a framework for retrieving three-dimensional facial structure, motion and spatio-temporal features from image sequences. First, we estimate monocular 3D scene flow by retrieving the facial structure using shape-from-shading (SFS) and combine it with 2D optical flow. Secondly, based on the retrieved structure and motion of the face, we extract spatio-temporal facial features for automated expression analysis. Experimental results illustrate the potential of the proposed 3D facial information retrieval framework for facial expression analysis, i.e. facial expression recognition and facial action-unit recognition on a benchmark dataset. This paves the way for future research on monocular 3D facial expression analysis.

I. INTRODUCTION

In the last two decades, automated analysis of social signals has attracted increased attention from researchers in neuroscience, psychology, cognitive sciences and computer science. Among the different types of social signals, particular attention has been paid to facial expression of emotions and its automated analysis from image sequences [1].

Automated facial expression analysis from image sequences is a very challenging task due to the complex three-dimensional deformation of the facial muscles, and the loss of 3D information during the image formation process. As a consequence, retrieving 3D spatio-temporal facial information from image sequences is essential for automated facial expression analysis [2], [3]. In this paper, we propose a framework for retrieving three-dimensional facial structure, motion and spatio-temporal features for automated expression analysis from image sequences.

First, we estimate monocular 3D scene flow by retrieving the face structure using shape-from-shading (SFS) and combine it with optical flow. For solving the SFS problem from a single image, we follow the method introduced in [4]. This method assumes Lambertian reflectance and uses spherical harmonics as approximation [5]. In this way, the SFS problem can be casted as an ill-posed linear least-squares problem which is regularized using a prior reference face shape and albedo map. Unlike in [4], we do not use a 3D face database to retrieve the prior reference face shape and albedo map. We instead fit, on the input image, the CANDIDE [6] 3D deformable face model. This to make our feature extraction framework robust against different face identities and facial expressions. We assume uniform albedo along the face and compute the albedo map by taking the average intensity value across the whole face. The obtained 3D face structure combined with 2D optical flow is then used to estimate the monocular three-dimensional scene flow. Considering an orthographic camera model, we solve the 3D scene flow estimation by following the methodology introduced in [7].

Secondly, based on the retrieved structure and motion of the face, we extract spatio-temporal facial features by leveraging the bag of visual words (BOVW) framework [8], [9]. The BOVW framework is a general pipeline to construct a global representation from a set of local features. As local features, we compute region covariance descriptors [10] of the retrieved 3D information around spatio-temporal interest points [11]. We then pre-process the obtained descriptors using the local...
log-Euclidean covariance matrix ($L^2$ECM) transform [12]. This allows transforming the local covariance matrices into vector-valued descriptors which can be handled with Euclidean operators. A codebook is then generated using the on-line dictionary learning algorithm described in [13]. Next, feature encoding is made via sparse coding (SC) [13], and finally a global representation is obtained using spatial pyramid pooling [14]. We illustrate the potential of our 3D facial information retrieval framework on two facial expression analysis tasks, namely, categorical facial emotion recognition, and facial action-unit recognition.

We summarize our main contributions as:

1. Robust facial shape-from-shading from monocular images using a 3D deformable face model;
2. Dense 3D scene flow estimation from monocular images using a facial shape and optical flow;
3. 3D spatio-temporal feature extraction for automated facial expression analysis.

The remainder of the paper is organized as follows. Section II summarizes the related work. Section III gives a detailed description of our framework. In Section III-A we develop our facial shape-from-shading and monocular 3D scene flow estimation method. Section III-B presents the spatio-temporal facial feature extraction framework. We then present our experimental results and findings in Section IV. Finally we conclude our work and discuss future work in Section V.

II. RELATED WORK

Within the past decades, significant research effort has been made to develop methods for facial expression analysis. However, most of the recent work on facial expression analysis is based on two-dimensional information [15], [16], [17]. Meanwhile, it is known that automated facial expression recognition from image sequences is a very challenging task due to the complex three-dimensional deformation and motion of the face associated to the facial expressions and the loss of 3D information during the image formation process. As a consequence, retrieving 3D spatio-temporal facial information from image sequences is essential for automated facial expression analysis.

With the rise of time-of-flight (ToF) cameras, most of the methods for 3D information retrieval rely on range data. Such data has recently been used for facial expression analysis. For a detailed overview of recent advances on facial analysis based on 3D information, we refer to [3]. Unlike these methods, we address the problem of 3D information retrieval from monocular brightness images. More specifically, we seek to retrieve the 3D facial shape and 3D facial scene flow. Most of the monocular 3D scene flow methods for facial analysis such as the ones proposed in [2] are based on the fundamental results introduced in [7]. In [2], the author used force models as regularization scheme for solving an ill-posed scene flow estimation problem. This requires a detailed description of a biomechanical model for obtaining accurate force models, hence accurate solutions [18]. More recently, a 3D facial analysis method based on hierarchical structure and non-rigid motion recovery has been proposed in [19]. The authors used a non-Lambertian shape-from-shading method [20] with predefined lighting conditions for recovering the facial structure and then used an affine motion model to estimate frame-by-frame motion parameters for facial expression recognition. Although their method doesn’t assume Lambertian surfaces, it relies on predefined lighting conditions, which is a very restrictive constraint. In [21], the authors presented a method for capturing facial details using space-time shape-from-shading. One major drawback of their method is the fact that they use stereo-vision for recovering the face shape, which seriously limits the scalability, hence applicability of the method. In our work, we use a monocular (single-view) approach.

The method presented in [22] leverages the large amounts of photos that are available per individual in personal or internet photo collections. In [23], the authors present a method that works without the need for any prior models or shape templates but has not been used in a recognition task. The method presented in [24] requires an external 3D facial expression database to construct the blend-shape mesh of the person. In [25], the authors present a method that requires a 3D face scan of the person for manual initialization. To our knowledge, we present one of the first works that achieve 3D facial information retrieval (structure and motion) from monocular brightness images without prior 3D information, training or manual initialization and uses it in a recognition task.

III. PROPOSED FRAMEWORK

Automatic facial expression analysis is aimed at objectively measuring the expression of emotions based on images. More precisely, the problem of automated facial expression analysis can be posed as the discovery of an unknown mechanism which establishes a mapping between the raw images and an output space of interest (e.g., basic emotions or facial action-units) through classification or regression. Frequently, this is not done directly on the input images, but on intermediate representation of images, which decomposes the problem into a two-stage processing: (i) feature extraction and (ii) classification/regression.

In this work we focus on the first processing stage, namely feature extraction, which consists of designing a mechanism that maps data from a raw image space to an intermediate image representation suitable for classification or regression. We propose a framework for retrieving three-dimensional facial structure, motion and spatio-temporal features for automated expression analysis from monocular image sequences. Our framework is composed of two major parts (see Figure 1). The first part retrieves three-dimensional facial information. This is achieved by fitting a 3D deformable model on the face image, followed by an estimation of the face structure using shape-from-shading (SFS) and combine it with 2D optical flow in order to obtain an estimate of the 3D scene flow. In the second part, spatio-temporal facial features are extracted based on the 3D face structure and motion retrieved in the first part. This is achieved by leveraging the bag of visual words (BOVW)
framework [8], [9], [26], which has become the most popular feature extraction framework during the last decade. The next sections discuss the two major steps of our framework, namely monocular 3D facial information retrieval and spatio-temporal feature extraction.

A. Monocular 3D Facial Information Retrieval

In this section we describe our method for retrieving 3D facial information from monocular image sequences. The specific problem we seek to solve, is the estimation of the 3D structure and motion of the face from its visible effects, namely the measured face image $I(x, y)$, defined on a compact domain $\Omega \subset \mathbb{R}^2$. In order to solve this problem, we proceed in three steps (see Figure 1). A first step consists of fitting a sparse 3D deformable face model $M_f$, called CANDIDE [6], on the face image $I$ and than refine the model using an interpolating subdivision scheme [27] in order to obtain a dense 3D deformable model $M'_f$ that we orthographically project onto the face image $I$, yielding a reference face shape $Z_{ref}(x, y)$. In a second step we use the obtained reference face shape as regularizer for solving an ill-posed shape-from-shading (SFS) problem and obtain a more accurate face shape $Z(x, y)$.

Fig. 1. Overview of proposed framework. The first group of processing blocks (left) illustrates the 3D information retrieval pipeline from raw images. The second group of processing blocks (right) illustrates the 3D spatio-temporal feature extraction pipeline (BOVW). The resulting image representation is further used for facial expression analysis.

Finally, in the third step, we use the obtained face shape together with 2D optical flow $\mathbf{v}(x, y)$ to estimate the 3D motion field $\mathbf{V}(x, y)$ in the image space $\Omega$, assuming an orthographic camera model.

1) 3D Deformable Face Model Fitting: The aim of this processing step is to obtain the reference face shape $Z_{ref}(x, y)$ representing the face structure in function of image coordinates. In a first stage we track 2D facial features points using the supervised descent method (SDM) [28]. More precisely, for a given image $I(x, y)$, we obtain $L$ facial feature points describing a 2D face shape $S_I = \{x_1, \ldots, x_L\} \subset \mathbb{R}^2$. In a second stage we use the tracked facial feature points $S_I$ for fitting the CANDIDE face model. The face model is represented by a set of pre-defined 3D control points $M_f = \{X_1, \ldots, X_C\} \subset \mathbb{R}^3$ connected by edges [6]. Each control point has its corresponding 2D point, yielding a pre-defined 2D face shape $S_{M_f} = \{x_1, \ldots, x_C\}$. Fitting the face model $M_f$ consists of deforming it such that its corresponding shape $S_{M_f}$ is aligned with the 2D face shape $S_I$. The 3D face model is then in its turn deformed according to the aligned 2D shape $S_{M_f}$ [29]. Once we obtain an aligned sparse 3D model $M_f$, we use the Catmull-Clark interpolating subdivision technique [30], [27] to obtain a dense face model $M'_f = \{X_1, \ldots, X_C\}$ with $C' \gg C$. As this dense model contains much more control points than pixels in the compact region it describes, we down-sample the dense model $M'_f$ and orthographically project it onto the image in order to obtain a reference face shape $Z_{ref}(x, y)$, densely defined in each pixel.

2) Facial Shape-from-Shading: To obtain a more accurate face shape, we use the obtained reference face shape $Z_{ref}(x, y)$ as regularizer for solving an ill-posed shape-from-shading (SFS) problem using the method introduced in [4]. Intuitively, this method uses the input image $I$ as a guide to “mold” the reference model in order to reach a desired reconstruction. Unlike in [4], we don’t use a 3D face database to retrieve the prior reference face model and albedo map. We instead estimate the reference face shape $Z_{ref}(x, y)$ as described above, which makes our framework more robust against different face identities and facial expressions. We construct an albedo map $\rho_{ref}(x, y)$ by assuming uniform albedo along the face and computing the albedo value by taking the average intensity value across the whole face. Following [4], assuming a Lambertian surface with albedo $\rho(x, y)$, the image irradiance equation equation can be expressed as

$$I(x, y) \approx \rho(x, y)I^T Y(n(x, y)),$$

and where $\rho = \frac{\partial Z}{\partial N}$, $q = \frac{\partial Z}{\partial N}$ and $N = \frac{1}{\sqrt{p^2 + q^2 + 1}}$. Consequently the unknown lighting $I$ is a four-dimensional vector. It has been shown that this approximation is highly accurate already when a low order (first or second) harmonic approximation is used [5]. Furthermore, modelling the reflectance with a spherical harmonic approximation allows the algorithm to handle multiple unknown light sources and attached shadows. Let $\mathbf{n}_{ref}(x, y)$ denote the normal to the reference face shape $Z_{ref}(x, y)$, the SFS problem can be posed as

$$\min_{\rho, Z} \int_{\Omega} (I - \rho I^T Y(n))^2 + \lambda_1 (\Delta G * d\rho)^2 + \lambda_2 (\Delta G * d\rho)^2 dxdy$$

(1)
where \( d_Z(x, y) = Z(x, y) - Z_{ref}(x, y) \) and \( d_\rho(x, y) = \rho(x, y) - \rho_{ref}(x, y) \). \( \Delta G^+ \) denotes convolution with the Laplacian of a Gaussian for smoothing the differences \( d_Z \) and \( d_\rho \), and \( \lambda_1, \lambda_2 \) are regularization parameters. The optimization problem (1) is approached by solving for lighting \( I \), depth \( Z \) and albedo \( \rho \) separately. For the lighting, in Equation (1), \( \rho \) is substituted by \( \rho_{ref} \) and \( Z \) by \( Z_{ref} \) (and consequently \( n \) by \( n_{ref} \)), which results in a over-constrained linear least squares optimization problem with only four unknowns and can be solved using the Moore-Penrose generalized inverse [31]. A similar approach is adopted for the depth \( Z \) by substituting \( I \) and \( \rho \) by the estimated lighting and the reference albedo \( \rho_{ref} \) respectively. This results in a large and sparse linear least squares optimization problem which can be solved using an algorithm such as the LSQR [32]. In this way, we obtain a more accurate estimate of the sought face shape \( Z(x, y) \) and further use it for the 3D scene flow estimation (see Figure 2).

3) 3D Scene Flow Estimation: Scene flow is an instantaneous 3D velocity field that describes the motion of a deformable surface imaged by a sensor [7]. It is defined as the 3D motion field \( V(x, y) \) of the points \( X = (X, Y, Z) \) in \( \mathbb{R}^3 \), just as optical flow is the 2D motion field \( v \) of the pixels \( x = (x, y) \) in an image \( I \) defined on a compact domain \( \Omega \subset \mathbb{R}^2 \). Our aim is to estimate the 3D motion field \( V(x, y) \) from the input image \( I(x, y) \) and its corresponding shape \( Z(x, y) \), obtained using the SFS method described above. We address this problem by following the methodology introduced in [7] and considering an orthographic camera model.

We first define the 3D motion field \( V \) as follows [7]:

\[
V \triangleq \frac{dX}{dt} = \frac{\partial X}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial X}{\partial t} \quad (2)
\]

where the first term is the projection of the scene flow on a plane passing through \( X \) tangent to the sensor’s projection ray, and the second term is the rate of change of the depth of the surface along the projection ray [2]. Then, assuming an orthographic camera model, the first and second terms of (2) can be computed as

\[
\frac{\partial X}{\partial t} \cdot \frac{dx}{dt} = \begin{bmatrix} \frac{\partial X}{\partial x} & \frac{\partial X}{\partial y} & \frac{\partial X}{\partial z} \end{bmatrix} \cdot \begin{bmatrix} dx \ dt \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} dx \ dt \end{bmatrix} \quad (3)
\]

and

\[
\frac{\partial X}{\partial t} = \begin{bmatrix} \|r(0)\|.r(x) & \|r(0)\|.r(z) \end{bmatrix} \cdot \frac{\partial Z}{\partial t} = \begin{bmatrix} 0 & 0 \\ \|r(0)\|.r(x) \end{bmatrix} \cdot \begin{bmatrix} dx \ dt \end{bmatrix} \quad (4)
\]

where \( r(0) \) is the unit vector in the direction of the camera optical axis and \( r(x) \) is the unit vector in the direction of the projection ray through pixel \( x \) [7]. For an orthographic camera model, both unit vectors are aligned with the z-axis, hence the simplification in (4).

By substituting the first and second terms of (2) by equations (3) and (4) respectively, we can express the 3D motion field as \( V = \left[ \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right]^T \), where \( \frac{dx}{dt} \) and \( \frac{dy}{dt} \) are the \( x \) and \( y \) components of the 2D optical flow. To compute the \( z \) component, we propose an implicit time-dependent formulation of the face shape \( Z(x(t), y(t)) \). In this way, using the chain-rule, we can express the \( z \)-component of the 3D motion field \( V \) as

\[
\frac{\partial Z}{\partial t} = \frac{\partial Z}{\partial x} \frac{dx}{dt} + \frac{\partial Z}{\partial y} \frac{dy}{dt} \quad (5)
\]

which is a function of the 2D optical flow and the spatial derivatives of the face shape \( Z(x, y) \) obtained using SFS. Figure 2 illustrates the obtained 3D information from monocular sequences. We can clearly observe the detailed 3D information, i.e. a facial shape and 3D scene flow (due to muscles deformations).

B. Spatio-Temporal Feature Extraction

Based on the retrieved 3D facial structure \( Z(x, y) \) and motion \( V(x, y) \), using the proposed methods described in previous sections (see Figure 2), we extract spatio-temporal facial features by leveraging the bag of visual words (BOVW) framework [8], [9]. Recently, BOVW has become the most popular feature extraction framework for constructing a global representation from a set of local features. BOVW is mainly composed of five steps (see Figure 1) (i) local feature extraction; (ii) feature pre-processing; (iii) codebook generation; (iv) feature encoding; (v) global feature pooling and normalization [9]. Among the several BOVW based feature extraction approaches that exist, very few are based on dynamic and three-dimensional information [9]. To address this issue, we apply BOVW feature extraction on top of the retrieved 3D facial information (see Figure 1). In the following, we give a description of the technical choices made in the proposed framework.
1) Local Feature Extraction and Pre-processing: As local features, we compute region covariance descriptors [10] to characterize local properties. This is achieved by extracting image patches around detected spatio-temporal interest points [11] and computing the sample covariance. Similar to histograms, covariance matrices are used to estimate the feature distribution in an extracted region [10]. However, compared to histograms, covariance matrices scale much better when the feature dimension grows. Given an image patch $R_k \subset \Omega$ of $N$ pixels around a spatio-temporal interest point $x_k \in \Omega$, the region covariance matrix is defined as

$$C_k = \frac{1}{N} \sum_{n=1}^{N} (f_k^{(n)} - \mu_k)(f_k^{(n)} - \mu_k)^T$$

where $f_k^{(n)}$ are the local image features and $\mu_k$ is the arithmetic mean of the features computed in region $R_k$. Note that in practice, we use a fast way for calculating covariance matrices based on the integral images, which makes the computation time independent of the region size [10]. In our framework, the local features $f_k^{(n)}$ can be 3D flow vectors $V(x, y)$, 3D normal vectors $N(x, y)$ [33] computed on the face shape $Z(x, y)$, or a 6D concatenation of both. They are referred to as covariance of flow (COF), covariance of normals (CON) and covariance of normals-flow (CONF).

Covariance matrices are symmetric and positive definite matrices (SPD), the space of which is a smooth Riemannian manifold [12]. Meanwhile, most of the classical machine learning algorithms assume vector-valued data coming from an Euclidean space. To this end, we further process the obtained covariance descriptors using the local log-Euclidean covariance matrix (L2ECM) transform [12] in order to obtain vector-valued descriptors which can be handled with Euclidean operators.

2) Codebook Generation and Feature Encoding: Generally, a codebook is generated by partitioning the descriptor space into codewords corresponding to regions represented by their center. Most of the recent work on facial expression analysis use k-means or dictionary learning for generating a codebook. In [15], the authors showed that both k-means and dictionary learning find the most frequent micro-temporal patterns in a given video sequence. They also observed that the dictionary learning algorithm introduced in [13] captures more salient properties of the local image descriptors than the k-means algorithm, and thus is more discriminative. Our framework, therefore uses dictionary learning (DL) on the L2ECM transformed covariance matrices as codebook generation method.

Encoding techniques can roughly be classified into three categories [9]: voting based encoding methods, super vector encoding methods and reconstruction based encoding methods. Given a codebook with $M$ words, $W = [w_1, \ldots, w_M] \in \mathbb{R}^{D \times M}$ with $D = \frac{d(d+1)}{2}$, the objective of encoding is to compute a representation $z_k \in \mathbb{R}^M$ for each input descriptor $c_k \in \mathbb{R}^D$ using codebook $W$. In voting based encoding such as hard vector quantization (VQ) [8], each descriptor directly votes for the “closest” codeword using a specific strategy, which causes much information loss [8]. Super vector encoding methods such as fisher vector (FV) [34] and vector of locally aggregated descriptors (VLAD) [35] generally yield high-dimensional representations by aggregating high order statistics, which make them less attractive for our feature extraction framework. In our framework, we choose to use reconstruction based encoding methods. Reconstruction based encoding methods are generative and seek to reconstruct the input descriptors using the codewords. The most popular reconstruction based encoding method is sparse coding (SC) and has been shown to yield good representations for facial expression analysis [15]. Our framework, therefore uses SC as encoding method.

3) Global Feature Pooling and Normalization: Given the encodings $z_k$ of all the local descriptors $c_k$, a pooling mechanism is used for obtaining a global representation $p \in \mathbb{R}^P$. For each image frame, depending on the number of detected spatio-temporal interest points, we may obtain a different number of local descriptors (and corresponding encodings). It is therefore essential to aggregate the local encodings into a vector of a fixed dimension, suitable as input representation to a machine learning algorithm for facial expression analysis.

Generally, techniques such as sum, average and max pooling are used as pooling mechanism [36] and have previously been used for facial expression analysis as e.g. in [15]. It is known that such pooling mechanisms disregard the spatial information of the used descriptors. We therefore propose to use spatial pyramid pooling (SPP) as mechanism. SPP partitions the image into $2^l \times 2^l$ increasingly finer sub-regions, typically with $l = 0, 1, 2$. In each sub-region, the local encodings are aggregated using max-pooling to form one intermediate representation $p_l$. The obtained representations of each sub-region are then concatenated to form a global representation of the observed image $p = [p_0^T, p_1^T, p_2^T]^T$. The global representation $p$ is further normalized such that it becomes invariant to the number of spatial sub-regions. In our framework, we use the $l_2$-normalization by dividing the $p$ by its $l_2$-norm: $p = p/\|p\|_2$, which yields the final (global) image representation suitable for facial expression analysis (see Figure 1).

IV. EXPERIMENTAL RESULTS

To evaluate our proposed framework, we perform facial expression recognition and facial action-unit recognition. Methods for facial expression analysis can be divided into frame-based and sequence-based methods [1]. The former typically employs static classifiers such as support vector machines (SVM). The latter employs dynamic classifiers such as dynamic Bayesian networks (DBN) and hidden Markov models (HMM). In our experiments, we take a holistic approach and pose the problem of sequence-based facial expression recognition as a multiple instance learning (MIL) problem [37]. We assume that each image sequence (bag of instances), associated to an expression or action-unit, contains at least one relevant frame while the remaining frames are possibly
irrelevant. As specific MIL model, we use multiple instance learning with embedded instance selection (MILES) [38].

To evaluate our framework (see Figure 1) combined with MILES classification, we use the Extended Cohn-Kanade (CK+) dataset [39], which is widely used for algorithm development and evaluation. We adopt a leave-one-subject-out cross-validation strategy for training and testing the MILES classifier on static (CON-BOVW), dynamic (COF-BOVW) and static-dynamic (CONF-BOVW) features. Furthermore, since the two recognition tasks are multi-class, we apply a simple one-versus-one heuristic with a voting scheme.

### TABLE I

<table>
<thead>
<tr>
<th>Features</th>
<th>Basic emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>An</td>
</tr>
<tr>
<td>CON-BOVW</td>
<td>91.0</td>
</tr>
<tr>
<td>COF-BOVW</td>
<td>89.8</td>
</tr>
<tr>
<td>CONF-BOVW</td>
<td>91.2</td>
</tr>
<tr>
<td>[39] SPTS</td>
<td>35.0</td>
</tr>
<tr>
<td>[39] CAPP</td>
<td>70.0</td>
</tr>
<tr>
<td>[39] SPTS+CAPP</td>
<td>75.0</td>
</tr>
</tbody>
</table>

For the facial expression recognition task, we trained a MILES classifier for automated recognition of 7 basic emotions: Anger (An), Contempt (Co), Happiness (Ha), Fear (Fe), Disgust (Di), Sadness (Sa) and Surprise (Su). Table I summarizes our obtained results. We also compare the performance of our framework, in terms of recognition rate, against the baseline results reported in [39]. Although they are slightly better, the baseline results are obtained using a frame-based classification strategy, while our results are obtained using a holistic (MIL) approach. The same holds for the comparison with the results reported in [16], which were obtained by only considering the last (apex) frame of each sequence.

From tables I and II, we observe that the proposed BOVW features based on our 3D facial information retrieval yield competitive results in terms of recognition rate. This confirms the potential of our framework for 3D facial expression analysis. Furthermore, we clearly observe the added value of 3D dynamic (COF and CONF) feature extraction compared to static (CON) feature extraction.

### V. CONCLUSIONS AND FUTURE WORK

We proposed a framework for retrieving three-dimensional facial structure, motion and spatio-temporal features for automated expression analysis from image sequences. We addressed the problem in two major steps. First, we estimated monocular three-dimensional scene flow. We achieved this by fitting a 3D deformable model on the face image, followed by a retrieval of the face structure using shape-from-shading (SFS) and combine it with optical flow. Secondly we extracted spatio-temporal facial features based on the retrieved 3D face structure and motion following the bag of visual words (BOVW) framework. We presented three major contributions: (1) robust facial shape-from-shading from monocular images using a 3D deformable face model; (2) dense 3D scene flow estimation from monocular images using a facial shape and optical flow; (3) 3D spatio-temporal feature extraction for automated facial expression analysis. Experimental results illustrate the potential of our framework in terms of facial expression recognition and facial action-unit recognition on a benchmark dataset. This paves the way for future research on monocular 3D facial expression analysis. As future work, we plan to further explore the BOVW framework for designing more informative features based on 3D facial information. We also plan to conduct experiments on more challenging datasets containing spontaneous facial expressions, as well as extending the methodology to dimensional affect recognition.

### ACKNOWLEDGMENT

This work is supported by the Agency for Innovation by Science and Technology in Flanders (IWT) – PhD grant nr. 131814, the VUB Interdisciplinary Research Program through the EMO-App project and the National Natural Science Foundation of China (grant 61273265).

### REFERENCES


