COMBINING AN ACTIVE SHAPE AND MOTION MODELS FOR OBJECT SEGMENTATION IN IMAGE SEQUENCES

Carlos Santiago, Jacinto C. Nascimento, Jorge S. Marques

Institute for Systems and Robotics (ISR/IST), LARSyS, Instituto Superior Técnico, Universidade Lisboa, Portugal

ABSTRACT

Obtaining the segmentation of an object in a sequence of images is usually achieved using a tracking methodology. However, in some applications, the whole sequence is available beforehand. This means that the segmentations can be determined simultaneously for all the frames in the sequence and taking into account the motion of the object. This paper proposes a new framework to incorporate motion information in the segmentation of image sequences using an active shape model (ASM). The motion of the object is modeled using a vector field, which is learned and refined online as the segmentation algorithm proceeds. The vector field is determined from the trajectories described by ASM points throughout the sequence. The vector field, in turn, influences the estimation of the ASM parameters by acting as a regularizer, ensuring that the segmentations are in agreement with the expected motion. The results show that coupling these models during the segmentation leads to an increase in performance, in particular by guarantee more consistent segmentations and by avoid gross errors in more challenging frames.

Index Terms— Segmentation, Active Shape Model, Motion Model, Vector Field

1. INTRODUCTION

Taking into account the motion of an object is often beneficial to achieve better results in its segmentation in sequences of images. The traditional approach is to adopt a tracking methodology, *e.g.*, the Kalman filter, in which the segmentation in a specific frame is influenced by the segmentation obtained in the previous frame, through the dynamical model. In some applications, however, the sequence is fully available beforehand. This means that the segmentations do not have to be determined sequentially, but can be estimated simultaneously. The advantage is that this allows each frame to influence (and be influenced by) all the remaining frames, and not just the following frame. This has the potential to improve the results by making the segmentations more robust in frames where the image conditions are more challenging.

An example where this approach is useful is the segmentation of cardiac magnetic resonance (CMR) data. A CMR sequence typically covers one cardiac cycle and the segmentation of the frames in which the heart is contracted is harder because boundaries are not so clearly defined [1]. Therefore, motion information allows the segmentation of these frames to be less dependent on image information and more reliant on the segmentations of the remaining frames.

In this work, we propose a framework to combine an active shape model (ASM) [2] with a motion model to determine the object position in the whole sequence simultaneously. This motion model acts as a regularizer in the estimation of the ASM parameters, ensuring that the segmentations are in agreement with the expected dynamics. Here, the motion model is described by a vector field that is learned and refined online, as the segmentation algorithm iterates, by analyzing the trajectories of the model points throughout the sequence. This information is then used to update the segmentations in a alternating scheme.

2. RELATED WORK

Adding motion information to an ASM-based methodology has been accomplished using two approaches: i) *embedding* this information within the deformation modes of a high-dimensional shape model [3, 4, 5, 6]; and ii) *combining* the ASM with a specific motion model [7, 8, 9, 10, 11, 12].

In the first approach (*embedding*), the traditional ASM is extended to jointly model the position of an object in an image sequence (instead of a single image). Similarly to the principal component analysis (PCA) performed to extract the deformation modes, in the extended version, the same analysis is used to extract not only local shape variations but also the variation along time. The downside of this approach is that it requires a large amount of training data, a problem that is typically called the *curse of dimensionality* [6, 13]. To overcome this limitation, most works resort to hierarchical ASMs [3, 5, 6, 9]. This variation of the traditional ASM divides the model into several patches and learns the shape statistics independently for each patch. This leads to a significant reduction

This work was supported by the FCT PhD grant [SFRH/BD/87347/2012], and FCT project and plurianual funding: [PTDC/EEIPRO/0426/2014], [UID/EEA/50009/2013].

in the data dimensionality. However, splitting the model into patches induces a loss of notion about the connectivity between patches, which may cause unexpected segmentations to be obtained. Furthermore, the model is also unable to capture the patterns of variation along time that the patches might share.

Regarding the second type of approach (*combining*), several motion models have been proposed to represent the object dynamics. A popular approach is to model the motion between consecutive frames using optical flow or registrations techniques [10, 11, 12]. The above methods provide a vector field between consecutive frames based on the comparison of intensity images (or other features derived from them). This type of approach is similar to the one proposed in this work. However, instead of using image features to determine the vector fields, we rely on the trajectories described by the estimated position of the ASM points along time. The following section describes the proposed methodology in more detail.

3. PROPOSED APPROACH

The proposed framework consists of alternating between two main steps: 1) **Estimating the motion model** based on the estimated ASM position in all frames; 2) **Updating the shape model parameters** based on observation points extracted from the image and according to the motion model estimated in 1). These steps are described next.

3.1. Estimating the motion model

We represent the motion model as a vector field (VF), defined on a regular grid within the image domain, where each vector in the grid determines the motion in a specific position of the image [14]. Let $\boldsymbol{V} = \left[\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_N^\top\right]^\top \in \mathbb{R}^{2N}$ denote the collection of the all the motion vectors in the grid with N nodes, where $\boldsymbol{v}_n \in \mathbb{R}^2$ is the vector associated with the n-th node. We define $V : \mathbb{R}^2 \mapsto \mathbb{R}^2$ as a function that maps a position in the image to the corresponding motion vector. For a generic position $\boldsymbol{x} \in \mathbb{R}^2$ within the image domain, the corresponding motion vector, $V(\boldsymbol{x}) \in \mathbb{R}^2$, is given by

$$\begin{aligned} V(\boldsymbol{x}) &= \sum_{n=1}^{N} \phi_n(\boldsymbol{x}) \boldsymbol{v}_n \\ &= \boldsymbol{\Phi}(\boldsymbol{x}) \boldsymbol{V}, \end{aligned} \tag{1}$$

where $\boldsymbol{\Phi}(\boldsymbol{x}) \in \mathbb{R}^{2 \times 2N}$ is a sparse matrix that determines the contribution of each node in the grid to the computation of $V(\boldsymbol{x})$, such that

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \begin{bmatrix} \phi_1(\boldsymbol{x}) & 0 & \ldots & \phi_N(\boldsymbol{x}) & 0 \\ 0 & \phi_1(\boldsymbol{x}) & \ldots & 0 & \phi_N(\boldsymbol{x}) \end{bmatrix}. \tag{2}$$

We adopt a bilinear interpolation scheme, where only the four closest grid nodes contribute to the computation of the motion vector, *i.e.*, at most, only fours elements of $\{\phi_1(\boldsymbol{x}), \ldots, \phi_N(\boldsymbol{x})\}$ are non-zero, and they satisfy the constraints $0 \leq \boldsymbol{\phi}_n(\boldsymbol{x}) \leq 1$ and $\sum_{n=1}^{N} \boldsymbol{\phi}_n(\boldsymbol{x}) = 1$.

Following [14], the VF, $\boldsymbol{V}$, is estimated from the trajectories described by each model point in the sequence. Let us assume we are given a set of K independent trajectories, such that the k-th trajectory is given by $\mathcal{X}^k = \{\boldsymbol{x}^k(1), \ldots, \boldsymbol{x}^k(T)\}$, where $\boldsymbol{x}(f) \in \mathbb{R}^2$ denotes the position of a point at a specific frame $f \in \{1, \ldots, T\}$. [1] The VF describes the trajectory of a point through the following dynamical model

$$\boldsymbol{x}(f) = \boldsymbol{x}(f-1) + V\big(\boldsymbol{x}(f-1)\big) + \boldsymbol{w} \tag{3}$$

where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma \boldsymbol{I})$ is white noise with Gaussian distribution. Replacing (1) into (3) we obtain

$$x(f) = \boldsymbol{x}(f-1) + \boldsymbol{\Phi}\big(\boldsymbol{x}(f-1)\big)\boldsymbol{V} + \boldsymbol{w}, \tag{4}$$

The maximum posterior estimate of the VF is given by (see details in [14])

$$\begin{aligned} \boldsymbol{V}^\star = \arg\min_{\boldsymbol{V}} \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{f=2}^{T} \left\| \Delta \boldsymbol{x}^k(f) - \boldsymbol{\Phi}\big(\boldsymbol{x}^k(f-1)\big)\boldsymbol{V} \right\|^2 + \\ + \alpha \boldsymbol{V}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{V}, \end{aligned} \tag{5}$$

where $\Delta \boldsymbol{x}^k(f) = \boldsymbol{x}^k(f) - \boldsymbol{x}^k(f-1)$, and the second term, $\alpha \boldsymbol{V}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{V}$, is related to a Gaussian prior. This prior acts as a smoothness regularizer by penalizing large differences between the vectors of neighboring nodes, encoded in $\boldsymbol{\Lambda}$. The parameter α determines the strength of the prior.

The solution of (5), $\boldsymbol{V}^\star$, is obtained by computing the derivative of the objective function with respect to $\boldsymbol{V}$ and equating to zero, which leads to the following linear equation

$$\begin{aligned} \left(\alpha \boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{f=2}^{T} \boldsymbol{\Phi}^\top\big(\boldsymbol{x}^k(f-1)\big) \boldsymbol{\Phi}\big(\boldsymbol{x}^k(f-1)\big) \right) \boldsymbol{V}^\star = \\ = \frac{1}{\sigma^2} \sum_{k=1}^{K} \sum_{f=2}^{T} \boldsymbol{\Phi}^\top\big(\boldsymbol{x}^k(f-1)\big) \Delta \boldsymbol{x}^k(f). \end{aligned} \tag{6}$$

3.2. Updating the shape model parameters

The ASM models the position of the contour point by

$$\boldsymbol{x}^i(f) = \boldsymbol{A}(f)(\overline{\boldsymbol{x}}^i + \boldsymbol{D}^i \boldsymbol{b}(f)) + \boldsymbol{t}(f), \tag{7}$$

such that $\overline{\boldsymbol{x}}^i$ and $\boldsymbol{D}^i$ are the mean position and the deformation modes, respectively, learned from a training set. The segmentation of the object in a frame f is determined by the set of ASM parameters $\widehat{\boldsymbol{\Theta}}(f) = \{\boldsymbol{A}(f), \boldsymbol{t}(f), \boldsymbol{b}(f)\}$. The estimation of these parameters is accomplished using an extension of

[1] Without loss of generality, in the following equations the trajectories are all assumed to be sampled at the same instants and the same number of times.

the EM-RASM [15] (expectation-maximization robust ASM) formulation for sequences of images. Given a set of observation points, $\{\boldsymbol{y}^i(f)\}_{i=1,\dots,M}, f = 1,\dots,T$, extracted from the image sequence, the shape model parameters are updated iteratively with the EM algorithm. In the E-step, the weight of each observation point is computed as follows

$$w^i(f) \propto \mathcal{N}(\boldsymbol{y}^i(f); \widehat{\boldsymbol{x}}^i(f), \boldsymbol{\Sigma}^i), \tag{8}$$

where $\widehat{\boldsymbol{x}}^i(f)$ depends on the ASM parameters estimate in the current iteration t, $\widehat{\boldsymbol{\Theta}}_{(t)}(f)$. These weights correspond to the likelihood of that observation belonging to the target object. In the M-step, the ASM parameters in all the frame are updated by minimizing a weighted least squares fit between the observations and the corresponding model points,

$$\{\widehat{\boldsymbol{\Theta}}(1),\dots,\widehat{\boldsymbol{\Theta}}(T)\}_{(t+1)} = \\ \arg\min_{\boldsymbol{\Theta}(1),\dots,\boldsymbol{\Theta}(T)} \sum_{f=1}^{T}\sum_{i=1}^{M} w^i(f) \left\|\boldsymbol{x}^i(f) - \boldsymbol{y}^i(f)\right\|^2_{\boldsymbol{\Sigma}^i}, \tag{9}$$

where $\|\boldsymbol{v}\|^2_{\boldsymbol{\Sigma}} = \boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v}$, $\boldsymbol{\Sigma}^i$ is a diagonal covariance matrix associated with $\boldsymbol{x}^i(f)$, and the position of $\boldsymbol{x}^i(f)$

In order to combine this estimation with the motion model, an additional *VF* term is included. This new term penalizes large deviations between each point, $\boldsymbol{x}^i(f)$, and its expected position according to the VF, given by

$$\boldsymbol{y}^i_V(f) = \boldsymbol{x}^i(f-1) + V(\boldsymbol{x}^i(f-1)), \tag{10}$$

where $V(\boldsymbol{x}^i(f-1))$ is computed by (1). Formally, we add the following new term,

$$\sum_{f=2}^{T}\sum_{i=1}^{M} \left\|\boldsymbol{x}^i(f) - \boldsymbol{y}^i_V(f)\right\|^2_{\boldsymbol{\Sigma}^i}, \tag{11}$$

to the objective function in (9), leading to

$$\{\widehat{\boldsymbol{\Theta}}(1),\dots,\widehat{\boldsymbol{\Theta}}(T)\}_{(t+1)} = \\ = \arg\min_{\boldsymbol{\Theta}(1),\dots,\boldsymbol{\Theta}(T)} \sum_{f=1}^{T}\sum_{i=1}^{M} w^i(f) \left\|\boldsymbol{x}^i(f) - \boldsymbol{y}^i(f)\right\|^2_{\boldsymbol{\Sigma}^i} + \\ + \lambda_V \left\|\boldsymbol{x}^i(f) - \boldsymbol{y}^i_V(f)\right\|^2_{\boldsymbol{\Sigma}^i}, \tag{12}$$

where λ_V is a constant that determines the importance of the VF term, and, for the first frame, $\boldsymbol{y}^i_V(1) = \boldsymbol{x}^i(1)$ (*i.e.*, the motion model does not influence the first frame).

We impose that $\boldsymbol{y}^i_V(f-1)$, computed using (10), is fixed given the current ASM parameters. Therefore, it does not depend on the $\widehat{\boldsymbol{\Theta}}(f-1)$. In practice, this can be seen as considering new observation points given by the expected motion, which are all weighted by λ_V. This will introduce a bias in the estimation of the ASM parameters that will make the segmentations combine the information from the images with the expected motion.

The solution of (12) is approximated by first minimizing with respect to the transformation parameters, $\boldsymbol{a}(f), \boldsymbol{t}(f)$, $f = 1,\dots,T$, and then minimizing with respect to the deformation coefficients, $\boldsymbol{b}(f)$, $f = 1,\dots,T$ (see details in [15]).

4. EXPERIMENTAL SETUP

The proposed approach is evaluated on two problems: 1) the segmentation of the left ventricle in CMR; and 2) the segmentation of lips in face images. For the first problem, we use the CMR sequence dataset [6], which comprises 33 sequences of volumes of healthy and disease patients. Each sequence contains 20 volumes, covering the systole and diastole phases of the cardiac cycle. Each volume contains 5-10 slices, with a spacing of 6-13 mm. Each slice is a 256×256 image, with a resolution of 0.93-1.64 mm, with a total of 5011 images. For the second problem, we use 61 sequences of 490×640 face images from the "happy" expression of the Cohn-Kanade expression database [16], each with 10-45 frames, for a total of 1241 images. The ground truth (GT) segmentations of both datasets is also provided.

The ASM (mean shape and modes of deformation) is learned using a leave-one-sequence-out strategy, *i.e.*, to test on a specific sequence, the model is learned using the remaining sequences. The quantitative evaluation of the segmentations is performed using two metrics: (i) the Dice coefficient, d_{Dice}, and (ii) the average perpendicular distance, d_{AV}.

In order to evaluate the advantage of the proposed approach, we compare our results with the segmentations obtained without the information from the motion model, *i.e.*, each frame is analyzed independently and without any notion of temporal dependency.

5. RESULTS

5.1. LV segmentation in CMR sequences

In the case of the LV segmentation, two VFs have to be learned: one for the contraction phase and one for the dilation phase. To separate these two phases, the trajectories of the model points along the sequence are divided in two parts. The first set of trajectories, associated to the frames $f = 1,\dots,f_{\text{s}}$, are used to compute the VF of the contraction phase, whereas the trajectories from frames $f = f_{\text{s}},\dots,T$ are used to compute the VF of the dilation phase. The switching frame f_{s} is determined as the frame in which the LV area was the smallest (*i.e.*, corresponding to the end of the contraction phase). An example of the VFs obtained is shown in Fig. 1.

Table 1 shows the average accuracy across the entire dataset with both metrics. It is possible to see that there is an increase in the performance when using information from the motion model. In particular, this information makes the segmentation algorithm give more consistent results across time and avoid many of the gross errors obtained when analyzing

Table 1. Mean (std) accuracy on LV segmentation in CMR sequences from [6].

Motion Model	None	VF
d_{Dice}	84.2 (9.2)	85.6 (7.4)
d_{AV}	2.5 (1.5)	2.3 (1.2)

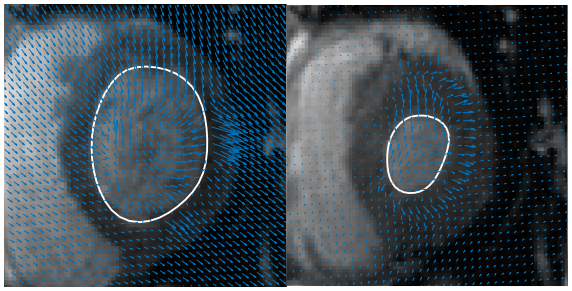

Fig. 1. Example of the VFs obtained to represent the motion of the LV during contraction (left) and dilation (right).

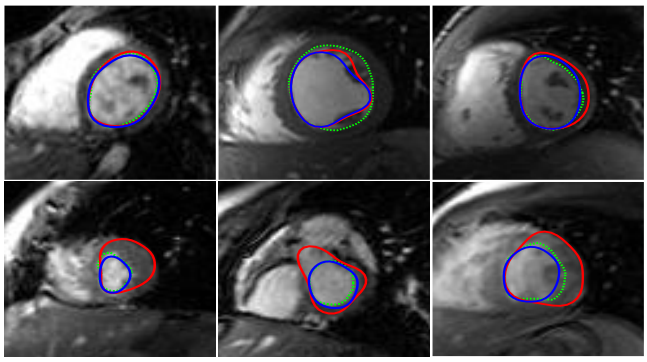

Fig. 2. Examples of LV segmentations with (blue) and without (red) the motion model and comparison with the GT (green).

each frame independently. Examples of the segmentations obtained with and without the motion model are shown in Fig. 2. In the examples shown in the top row, both methods have a similar performance. However, in the bottom row we show examples where there is a clear advantage in using the motion model.

5.2. Lip segmentation in face image sequences

In the lip segmentation problem, we observed the same advantages discussed above. The accuracy of the segmentation is increased when the motion model is included, leading to the overall improvement shown in Table 2. An example of the VF obtained in a particular sequence is shown in Fig. 3, where it is possible to see that it represents the motion of the lips when a person starts smiling. Fig. 4 shows a comparison between the segmentation obtained with and without the motion model. As previously, the top row shows examples in

Table 2. Mean (std) accuracy on lip segmentation in face image sequences from [16].

Motion Model	None	VF
d_{Dice}	85.1 (6.9)	86.2 (5.6)
d_{AV}	3.1 (1.4)	2.9 (1.1)

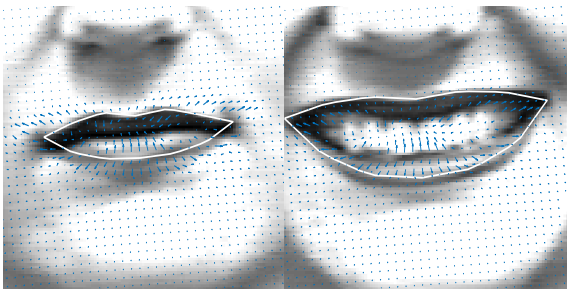

Fig. 3. Example of the VF obtained to represent the motion of the lips.

Fig. 4. Examples of lip segmentations with (blue) and without (red) the motion model and comparison with the GT (green).

which the performance is similar in both approaches, but the bottom row shows examples where it was beneficial to use the VF constraint.

6. CONCLUSIONS

This paper proposes a framework to incorporate motion information for the segmentation of sequences using an ASM-based model. In this new framework, the segmentation of each frame in the sequence occurs simultaneously. The motion of the object is modeled using a VF, which is learned and refined online, as the segmentation algorithm iterates. The VF acts as a regularizer in the estimation of the ASM parameters, ensuring that the segmentations obtained are in agreement with the expected motion of the object. The results show that there is a clear advantage in using this information during the segmentation, in particular to guarantee consistency between segmentations in different frames and by avoid gross errors in more challenging frames.

7. REFERENCES

[1] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M. Jolly, A. H. Kadish, et al., "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images," *Medical image analysis*, vol. 18, no. 1, pp. 50–62, 2014.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[3] C. Davatzikos, X. Tao, and D. Shen, "Hierarchical active shape models, using the wavelet transform," *IEEE transactions on medical imaging*, vol. 22, no. 3, pp. 414–423, 2003.

[4] G. Hamarneh and T. Gustavsson, "Deformable spatio-temporal shape models: extending active shape models to 2D+ time," *Image and Vision Computing*, vol. 22, no. 6, pp. 461–470, 2004.

[5] J. Lötjönen, K. Antila, E. Lamminmäki, J. Koikkalainen, M. Lilja, and T. Cootes, "Artificial enlargement of a training set for statistical shape models: Application to cardiac images," in *International Workshop on Functional Imaging and Modeling of the Heart*. Springer, 2005, pp. 92–101.

[6] A. Andreopoulos and J. K. Tsotsos, "Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI," *Medical Image Analysis*, vol. 12, no. 3, pp. 335–357, 2008.

[7] F. Billet, M. Sermesant, H. Delingette, and N. Ayache, "Cardiac motion recovery and boundary conditions estimation by coupling an electromechanical model and cine-MRI data," in *Functional Imaging and Modeling of the Heart*, pp. 376–385. Springer, 2009.

[8] C Casta, P Clarysse, J Schaerer, and J Pousin, "Evaluation of the dynamic deformable elastic template model for the segmentation of the heart in MRI sequences," *MIDAS J-Card MR Left Ventricle Segmentation Challenge*, 2009.

[9] S-W Lee, J Kang, J Shin, and J Paik, "Hierarchical active shape model with motion prediction for real-time tracking of non-rigid objects," *IET Computer Vision*, vol. 1, no. 1, pp. 17–24, 2007.

[10] L. A. Schwarz, A. Mkhitaryan, D. Mateus, and N. Navab, "Human skeleton tracking from depth data using geodesic distances and optical flow," *Image and Vision Computing*, vol. 30, no. 3, pp. 217–226, 2012.

[11] O. Arif, G. Sundaramoorthi, B. Hong, and A. Yezzi, "Tracking using motion estimation with physically motivated inter-region constraints," *IEEE transactions on medical imaging*, vol. 33, no. 9, pp. 1875–1889, 2014.

[12] L. Wang, A. Basarab, P. R. Girard, P. Croisille, P. Clarysse, and P. Delachartre, "Analytic signal phase-based myocardial motion estimation in tagged mri sequences by a bilinear model and motion compensation," *Medical image analysis*, vol. 24, no. 1, pp. 149–162, 2015.

[13] S. P. O'Brien, O. Ghita, and P. F. Whelan, "A Novel Model-Based 3D Time Left Ventricular Segmentation Technique," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 2, pp. 461–474, 2011.

[14] J.C. Nascimento, M.A.T. Figueiredo, and J.S. Marques, "Activity Recognition Using a Mixture of Vector Fields," *Image Processing, IEEE Transactions on*, vol. 22, no. 5, pp. 1712–1725, May 2013.

[15] Carlos Santiago, Jacinto C Nascimento, and Jorge S Marques, "2D Segmentation Using a Robust Active Shape Model With the EM Algorithm," *Image Processing, IEEE Transactions on*, vol. 24, no. 8, pp. 2592–2601, Aug 2015.

[16] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, 2010, pp. 94–101.