Semi-automatic Tracking of Markers in Facial Palsy

Philip Limbeck, Walter G. Kropatsch, and Yll Haxhimusa
Vienna University of Technology, Pattern Recognition and Image Processing Group
{*phil,krw,yll*}*@prip.tuwien.ac.at*

Abstract

We introduce a semi-automatic tracking method that can be utilized for the analysis of facial markers in the medical condition of facial palsy. Tracking of markers will help medical physicians in evaluating this medical condition quantitatively. We use particle filtering to track markers towards measuring distances needed to evaluate the degree of facial palsy. We show that by employing tracking methods, the analysis time is reduced without losing the high accuracy of the results.

1. Introduction

Facial palsy is a medical condition that leads to nerval degeneration and subsequently paralysis of different facial muscles. This often results in facial asymmetry especially around eyes and mouth (Fig. 1). Qualitative measures for facial palsy - such as medical rating scales - do not allow measuring the healing progress over time because of their high intra- and inter-observer variability. An alternative are quantitative methods which find the distances over time (tracking) of facial landmarks [6]. Each captured video of a patient contains several actions s/he has to perform needed to provide measurement data for quantifying the extent of facial palsy. The number of markers and their position in the face is given by the physicians at Vienna Medical University (VMU) [6]. 15 dynamic markers are placed on the patients dynamic part of the face, like mouth, eyes and nose; and 3 static markers of different color are attached to the rhinion and each tragus (Fig 1). The existing manual method in use at VMU takes up to 5 hours for the analysis of the patient's video [6]. Our goal is to reduce the video analysis time, keeping the result's accuracy by employing tracking methods with manual intervention support.

Many methods have been introduced [3, 5, 7] for capturing and analyzing facial expressions. Shu et al. [8, 9] used classification based on *local binary patterns* (*LBP*) [12] to evaluate the degree of palsy using qualitative measures. The disadvantage of this method is that the temporal information is lost. The palsy analysis would benefit from including the temporal information, as it is shown in this paper. For this medical application, the particle filter tracking method [1] showed the best results compared to other tracking methods like mean shift tracking [4] and optical flow and optical strain based tracking [2, 13]. The particle filter validates a set of hypothesis against image measurements. This validated set of hypothesis is used to estimate a multimodal probability distribution, which makes it robust against occlusion and clutter [16]. Our contribution in this paper is the exploration of the suitability of a particle filtering algorithm by applying different likelihood models to the problem of tracking landmarks for analysis of facial palsy. We show that semi-automatic tracking of the landmarks can reduce the analysis time. This paper is structured as follows. In Sec. 2 we shortly explain the chosen method. The experiments and results are given in Sec. 3, followed by the conclusion.

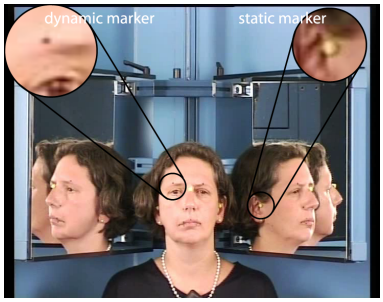

Figure 1: Dynamic and static facial markers and the two mirror capturing video equipment in usage at VMU.

2. Tracking with Particle Filtering

We build a set of the initial landmarks that are selected by the user. The coordinates of the landmarks along with a size of the markers are used to build a set of target models $\{A_k\}_{k=1}^{T}$. The posterior dis-

tribution $P(X_{t-1}|Z_{0:t-1})$ at time $t-1$ is the set of particles $\left\{X_{t-1}^{(i)}, w_{t-1}^{(i)}\right\}_{i=1}^{N}$. An initial set of particles $\left\{X_0^{(i)}, w_0^{(i)}\right\}_{i=1}^{N}$ is generated from the coordinates and subdivided into different unlabeled clusters corresponding to the initial locations, separating the different modes of the posterior distribution $P(X_{t-1}|Z_{0:t-1})$. Every image frame in the video is denoised using a bilateral filter [14]. Afterwards, we segment the face using thresholding in the YCbCr color space. Finally, morphological closing and removal of large segments produce an image which is fed into the particle filter. To associate the unlabeled clusters with the labeled targets, a cost matrix is created using the L^2 distances between the clusters and the labeled targets, which is fed into the Hungarian algorithm [10]. For state transition, we use a first order motion model $X_t^{(i)} = M_t^{(i)} X_t^{(i)} + w_t$ with M_t being the state transition matrix and $w_t \sim N(0, \sigma)$. The value of σ depends on the selected observation model. Systematic resampling is used according to the effective sample size $1/\sum_{i=1}^{N} |w_i|^2$. Each measurement Z_k of the available measurements $\{Z_k\}_{k=1}^{T}$ is associated with a specific target which in turn is associated with a cluster of particles. This partitioning of the state space allows us to track multiple targets using a single particle filter. To restrict the search space of the markers, we employed Voronoi tesselation in the observation model. Details on the tracking model are found in [15]. In this paper we study two different likelihood models for tracking:

Template-Based Particle Filter (TPF). For the template-based particle filter, the state X_t is defined as $\begin{bmatrix} x & \dot{x} & y & \dot{y} \end{bmatrix}$. The appearance model is based on the intensity templates and LBP to incorporate texture information. Both information is combined to a single likelihood image $L_k = (1-\beta)L_{color} + \beta L_{LBP}$. The value of β controls the influence of the texture information towards the likelihood. To explore this influence, apart from the base method with β set to 0, two other methods are derived. For the Mixed-TPF the value is set to 0.2 and for the filter which is only using LBP information (LBP-TPF), a value of 1.0 is used. The likelihood image L_k is computed in a search region surrounding the position of the landmark in the previous frame by applying normalized cross correlation over the current frame. The likelihood for a given sample $X_t^{(i)}$ at time t is computed by

$$P_t(Z_k|X_t^{(i)}) = 1/(\sqrt{2\pi}\sigma) e^{-\psi(X_t^{(i)})^2 \frac{1}{2*\sigma^2}}, \quad (1)$$

$$\psi(x) = 1 - (L_k(W(x)) + 1)/2, \quad (2)$$

where $W(x)$ warps the image coordinates into likelihood image coordinates. The appearance model for the target k is updated if the likelihood of the estimated position exceeds a threshold U_γ using: $A_k = (1-\gamma)A_k + \gamma\tilde{A}_k$, where $\tilde{A}_k$ is the candidate model of the current position estimate.

Color-Based Particle Filter (CPF). For the color-based particle filter the state $\begin{bmatrix} x & \dot{x} & y & \dot{y} & H_x & H_y & \alpha \end{bmatrix}$ is used, where H_x and H_y are the elliptic axis and α being the rotation. The Bhattacharyya coefficient $\rho[p_s(i), q]$ is used to calculate the congruency between the target and candidate histograms. The complete observation model has been adopted from [11] except that we changed the distance calculation to: $\sqrt{1-\rho[p_s(i), q]}$ [4], because this represents a proper distance bounded with $[0, 1]$.

3. Experimental Setup and Results

Video Sequences. Five subjects are video recorded performing several clinically relevant facial expressions (e.g. smiling) [6]. The first three subjects (S1-S3) are recorded using our camera setup, with 1044×1080 resolution, and the last two subjects (S4-S5) with the VMU setup (Fig.1), with 736×576 resolution. The ground truth (Tab. 1) consists of the positions obtained by manually locating the markers in each frame, and subsequently using the Hungarian algorithm to create a trajectory for each marker over time.

Parameter Selection. We applied the two different methods (TPF and CPF) to five video sequences of subjects (S1 to S5) with attached landmarks of different color. To evaluate the accuracy of the tracking method, the L^2 distances between the ground truth and the estimated coordinates are computed. To analyze the stability of the estimate, the weighted Average Absolute Deviation (AAD_w): $\frac{1}{n}\sum_{i=1}^{n} w_i |X^{(i)} - \hat{X}|$ is used. To allow comparison with manual method, the number of interventions and the total time to process the sequence is measured. After an extensive evaluation of the model parameters, we selected the following parameter values for the experiment described in this paper. The standard deviation σ of the likelihood model is set to 0.05. Since we expect more variance on the y than on the x image axis, the deviations in the corresponding direction, σ_x and σ_y are set to $\sqrt{0.5}$ and 1.0 respectively. Depending

Video	*#Frames*	*# Markers*	*Time (min)*
S1	623	15	**43**
S2	577	15	**31**
S3	446	15	**41**
S5	1510	34	**540**

Table 1: Time needed to manually locate the markers.

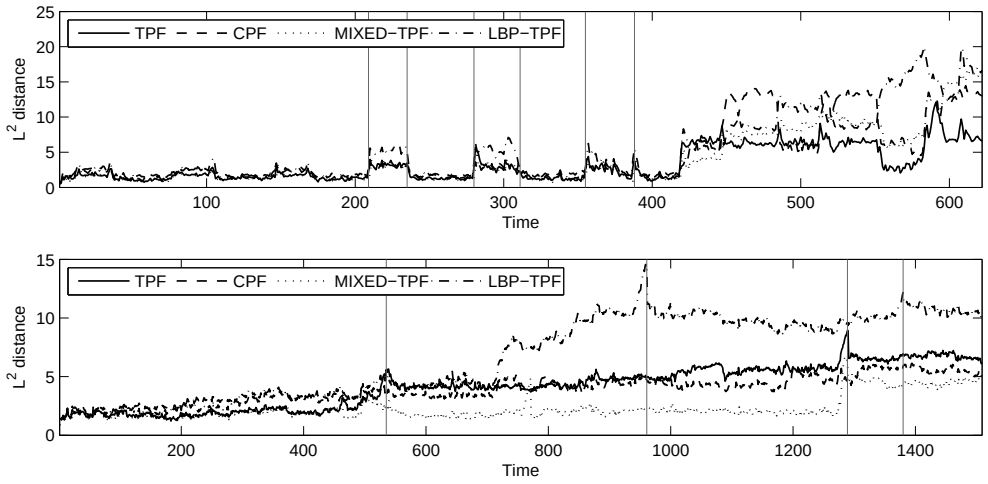

Figure 2: Performance of all particle filter methods for sequences S1 (above) and S5 (below).

Method	*Video*	*Avg. RMSE*	*Avg.* AAD_w	*Avg. Fps*	*# Interac.*	*Time (min)*
CBF	S1	9.92	0.036	0.14	0	**22**
	S2	13.14	0.036	0.16	2	**26**
	S3	13.54	0.036	0.16	0	**18**
	S5	6.67	0.024	0.09	2	**81**
TPF	S1	6.84	0.034	0.09	0	**14**
	S2	3.87	0.036	0.11	0	**16**
	S3	12.69	0.032	0.15	0	**17**
	S5	7.69	0.013	0.13	2	**107**
LBP-TPF	S1	12.26	0.029	0.22	0	**34**
	S2	10.48	0.032	0.22	0	**32**
	S3	14.47	0.024	0.23	0	**35**
	S5	12.45	0.016	0.17	5	**145**
Mixed-TBF	S1	9.33	0.031	0.52	0	**80**
	S2	4.77	0.031	0.28	0	**40**
	S3	11.96	0.029	0.37	0	**41**
	S5	4.84	0.012	0.15	13	**127**

Table 2: Evaluation results of different methods. Time needed to semi-automatically locate the markers.

on the dataset a different marker size has to be selected to ensure that the right region is tracked. A coarse-to-fine tracking method could be employed instead to omit the step of manually providing the marker size. For the color-based particle filter, σ_ρ has been set to 0.05 radiant. Additionally, the parameters for σ_{H_x} and σ_{H_y} have been set both to 0.1. We chose the state transition matrix M_t being a diagonal matrix with the values of Δ_{H_x} and Δ_{H_y}, the relative change of the size, set to 1.0. The relative positional change Δ, has been derived from the marker size, because although the marker size does not necessarily represent the pixel motion of the face, it still is a good indicator how large the expected motion will be. Hence, we set the value of $\Delta = 2 \cdot (30/size_y(A_k))$ by computing an aspect ratio between two sequences of the same marker size but - due to a different resolution - a different pixel size. The change of the marker rotation over time is only modeled as noise parameter and not reflected in M_t. Because we expect the features of having a low discriminative power, we chose a value of 250 for the number of particles. For the color-based particle filter, a weighted marginal histogram with a bin size of 8 for each channel is used. The HSV influence of each channel is given by [0.1, 0.6, 0.3] to suppress the hue fraction of the human face. The value of γ has been set to 0.2 and the threshold U_γ has been set to 4.8, which

corresponds to the selection of one standard deviation from the mean in the observation model.

Discussion. The results in Tab. 2 and Fig. 2 show the results of the evaluated video sequences (S1-S5). TPF performs best for every dataset except S3. This can be explained by fast appearance changes in S3, caused by contractions of the mouth region. (THIS IS NEW!) The values of Avg. AAD_w are all within an interval of $[0.01, 0.03]$ which can be explained by using the same values of σ. Although the low value of σ increases the RMSE on the sequence, it also increases the risk for a target loss since particles with a lower likelihood will not survive and regain the target in case of fast movement. In sequence S5, one marker is lost very early because it is located right next to the facial border which results in removal due to facial segmentation. This might be prevented using a more sophisticated method of extracting the facial contour. Additionally, the threshold for model updates U_γ has to be set very carefully. If the value is set too low, all TPF-based trackers lose their targets within 100-200 frames. If this value is set too high, the tracker tends to drift away because background information is falsely incorporated into the model. The performance can be increased even by using more particles, on the cost of the tracking time. Since the features are not discriminative enough in some regions of the face, even multiple hypothesis might result in the particle filter being trapped in a local maximum which is similar to the appearance model. This is not an issue with CPF, since color (green, blue, and orange) markers have been selected to be most dissimilar from the face. Although tracking happens automatically, in some cases the position has to be corrected manually to ensure that accuracy is low enough for clinical usage. Fig. 2 show the average L^2 distances for subjects S1 and S5. The peaks (vertical lines in Fig. 2) denote where the raising eyebrows expression starts and ends in S1. It can be seen that they coincide with a performance loss, i.e. the 4 markers around the eye are lost by all trackers except TPF. After the peak expression is over, the markers are redetected and tracking continues. The performance degrades over time, and can be explained by sample impoverishment due to resampling, as well as weak features that can not discriminate the markers from the background.

4. Conclusion

We introduced a semi-automatic tracking method, based on the particle filter as the first step needed by the medical physicians to diagnose the degree of the facial palsy. Measuring marker's distances and tracking them over the time allows to measure quantitatively the degree of facial palsy. We show that by using tracking method the analysis time is reduced. It has been also shown that the accuracy results are acceptable, but further investigation are needed to achieve the high degree of accuracy required by the clinical applications.

References

[1] M. S. Arulampalam, S. Maskell, and N. Gordon. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing*, 50:174–188, 2002.

[2] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. on PAMI*, 33(3):500 –513, 3 2011.

[3] B. F. Christian and C. Küblbeck. Face tracking by means of continuous detection. In *Proc. of CVPR Workshop on Face Processing in Video*, 2004.

[4] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. of CVPR*, 142–149, 2000.

[5] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *J. of PR*, 36(1):259–275, 2003.

[6] M. Frey, P. Giovanoli, H. Gerber, M. Slameczka, and E. Stussi. Three-dimensional video analysis of facial movements: A new method to assess the quantity and quality of the smile. *Plastic and Reconstructive Surgery*, 104(7):2032–2039, 1999.

[7] H. Gu, Q. Ji, and Z. Zhu. Active facial tracking for fatigue detection. In *Proc. of 6th IEEE Workshop on Applications of Computer Vision*, 137–142, 2002.

[8] S. He, J. Soraghan, B. O'Reilly, and D. Xing. Quantitative analysis of facial paralysis using local binary patterns in biomedical videos. *IEEE Trans. on Biomed. Engineering*, 56(7):1864–1870, 7 2009.

[9] S. He, J. J. Soraghan, and B. F. O'Reilly. Biomedical image sequence analysis with application to automatic quantitative assessment of facial paralysis. *J. Image Video Process.*, (2):1–11, 2007.

[10] H. W. Kuhn. The hungarian method for the assignment problem. In *NRLQ*, 2(1-2):83–97, 1955.

[11] K. Nummiaro, E. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. J. of Image and Vision Comp. (21)1:99110, 2002.

[12] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *J. of PR*, 29(1):51 – 59, 1996.

[13] M. Shreve, N. Jain, D. Goldgof, S. Sarkar, W. G. Kropatsch, C.-H. J. Tzou, and M. Frey Evaluation of facial reconstructive surgery on patients with facial palsy using optical strain. In *CAIP*, 512-519, 2011.

[14] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 839 –846, 1998.

[15] P. Limbeck. Interactive Tracking of Markers for Facial Palsy Analysis. Diploma Thesis, Vienna University of Technology, 2012.

[16] E. Maggio and A. Cavallaro. Video Tracking, Theory and Practise. Wiley, 98–111, 2011.