Local Feature Evaluation for a Constrained Local Model Framework

Maiya Hori(✉), Shogo Kawai, Hiroki Yoshimura, and Yoshio Iwai

Graduate School of Engineering, Tottori University,
101 Minami 4-chome, Koyama-cho, Tottori 680-8550, Japan
hori@ike.tottori-u.ac.jp

Abstract. We present local feature evaluation for a constrained local model (CLM) framework. We target facial images captured by a mobile camera such as a smartphone. When recognizing facial images captured by a mobile camera, changes in lighting conditions and image degradation from motion blur are considerable problems. CLM is effective for recognizing a facial expression because partial occlusions can be handled easily. In the CLM framework, the optimization strategy is local expert-based deformable model fitting. The likelihood of alignment at a particular landmark location is acquired beforehand using the local features of a large number of images and is used for estimating model parameters. In this learning phase, the features and classifiers used have a great influence on the accuracy of estimation in landmark locations. In our study, tracking accuracy can be improved by changing the features and classifiers for parts of the face. In the experiments, the likelihood map was generated using various features and classifiers, and the accuracy of landmark locations was compared with the conventional method.

1 Introduction

In recent years, communication robots [1] used in applications such as guidance and nursing care have been developed. These communication robots have a camera and a microphone and can communicate with the target by recognizing speech and facial expressions. Figure 1 shows a cellphone-type tele-operated communication medium called Elfoid [2]. Elfoid is designed to transmit the speaker's presence to the communication partner using a camera and microphone. When using this type of robot for communication, it is important to convey the facial expressions of the speaker to increase communication modality. If the speaker's facial movements can be regenerated accurately using these robots, the human presence can be adequately conveyed. Elfoid has a camera within its body and the speaker's facial movements can be estimated through an accurate facial recognition approach. When we recognize facial images captured by a mobile camera such as Elfoid, changes in lighting conditions and image degradation from motion blur are considerable problems. In this study, we aim to construct a face tracking technique that works robustly even under severe conditions.

Q. Ji et al. (Eds.): FFER 2014, LNCS 8912, pp. 11–19, 2015.
DOI: 10.1007/978-3-319-13737-7_2

Fig. 1. Cellphone-type tele-operated android: Elfoid. It is important to convey the facial expressions of the speaker to increase communication modality.

2 Related Work

Face tracking techniques using feature points such as the corners of the eyes and mouth are effective for recognition of facial expressions because a face is a non-rigid object. Deformable model fitting approaches [3][4] have been proposed. Facial deformable models can be divided into two main categories, holistic and part based models.

A notable example of a holistic model is the Active Appearance model (AAM) [3]. Holistic models employ a Point Distribution Model (PDM) as:

$$\mathbf{x_i} = s\mathbf{R}(\overline{\mathbf{x_i}} + \Phi_i \mathbf{q}) + \mathbf{t}, \tag{1}$$

where $\mathbf{x_i}$ denotes the 2D location of the PDM's ith landmark, s denotes global scaling, $\mathbf{R}$ denotes a rotation and $\mathbf{t}$ denotes a translation. $\overline{\mathbf{x_i}}$ denotes the mean location of the ith PDM landmark in the reference frame and Φ_i denotes the basis of the variations. $\mathbf{q}$ is a set of non-rigid parameters. A statistically shaped model is acquired from a set of training points by applying principal component analysis (PCA). The algorithm uses the difference between the current estimate of the shape model and the target image, to drive an optimization process. However, holistic approaches have many drawbacks. They are sensitive to lighting changes, and partial occlusions cannot be easily handled.

Part-based models use local image patches around the landmark points. Constrained Local Models[4] outperform AAM in terms of landmark localization accuracy. CLM fitting is generally posed as a search for the PDM parameters, $\mathbf{p}$, that minimize the misalignment error in following Eq. (2):

$$\mathcal{Q}(\mathbf{p}) = \mathcal{R}(\mathbf{p}) + \sum_{i=1}^{n} \mathcal{D}_i(\mathbf{x}_i; \mathcal{I}), \tag{2}$$

where $\mathcal{R}$ is a regularization term and $\mathcal{D}_i$ denotes the measure of misalignment for the ith landmark at $\mathbf{x}_i$ in the image $\mathcal{I}$. In the CLM framework the objective is to create a shape model from the parameters $\mathbf{p}$. In [4], the regularization and misalignment error function in Eq. (2) take the following forms:

$$\mathcal{R}(\mathbf{p}) = -\ln p(\mathbf{p}), \tag{3}$$

$$\mathcal{D}_i(\mathbf{x}_i; \mathcal{I}) = -\ln p(l_i = 1 | \mathbf{x_i}, \mathcal{I}). \tag{4}$$

CLM models the likelihood of alignment at a particular landmark location, $\mathbf{x}$, as follows:

$$\mathbf{p}(l_i = 1 | \mathbf{x}, \mathcal{I}) = \frac{1}{1 + \exp\{l_i \mathcal{C}_i(\mathbf{x}; \mathcal{I})\}}, \tag{5}$$

where $\mathcal{C}_i$ denotes a classifier that discriminates aligned from misaligned locations. The likelihood of alignment at a particular landmark location is acquired beforehand using local features of a large number of images. To generate the classifier $\mathcal{C}_i$, Saragih et al. [4] use logistic regression. Mean-shift vectors from each landmark are computed using the likelihood of alignment and the parameters $\mathbf{p}$ are updated. These processes are iterated until the parameters $\mathbf{p}$ converge.

When we recognize facial images captured by a mobile camera within Elfoid, changes in lighting conditions and image degradation from motion blur are considerable problems. In [4], the changes in the environment are not specifically considered when calculating the likelihood of alignment. Furthermore all landmarks are treated in the same manner. In our study, it is intended to construct a face tracking technique that works robustly even under severe conditions. To adapt to the changes in environment, the likelihood of alignment is calculated using various features that are robust for a particular situation. The PDM parameters, $\mathbf{p}$ in Eq. (2) are updated in the same manner as the conventional technique[4] using the estimated likelihood of alignment. Conclusively, tracking accuracy can be improved by changing the features and classifiers for different parts of the face.

3 Evaluation of the Likelihood of Alignment at a Particular Landmark Location

In this study, landmark locations are estimated using the likelihood maps and the accuracy of their positions is evaluated. The likelihood maps at each landmark location are generated according to Eq. (5). In Eq. (5) the classifier is generated using image features extracted at manually annotated landmark locations. The details of the local features and the classifier are described in the following sections.

3.1 Local Features

As features, the gray scaled patch, image gradient, Local Binary Patterns (LBP) [5], Local Directional Pattern (LDP) [6], Local Phase Quantization (LPQ) [7],

SIFT [8], SURF [9] and HOG [10] are used for producing response maps. The most important property of the LBP operator is its robustness to gray-scale changes caused by illumination variation. A LDP feature is obtained by computing the edge response values in all eight directions at each pixel position and generating a code from the magnitude of the relative strength. LPQ is based on quantizing the Fourier transform phase in the local neighborhood and is robust against the most common image blurs.

3.2 Classifiers

The positions of landmarks are estimated in the likelihood map as shown in Fig. 2. The likelihood in Eq. (5) is generated using a classifier $\mathcal{C}_i$. Classifiers are generated using logistic regression and support vector machines (SVMs). Logistic regression is a type of probabilistic statistical classification model. An SVM model is a representation of the examples of points in space, mapped so that the examples of the categories are separated by a clear gap that is as wide as possible.

3.3 Facial Image Database Captured by a Mobile Camera

Facial images used in this study are captured by a camera embedded in Elfoid. This database includes images that have various facial expressions, lighting changes and partial occlusions. The subject changes its head pose considerably

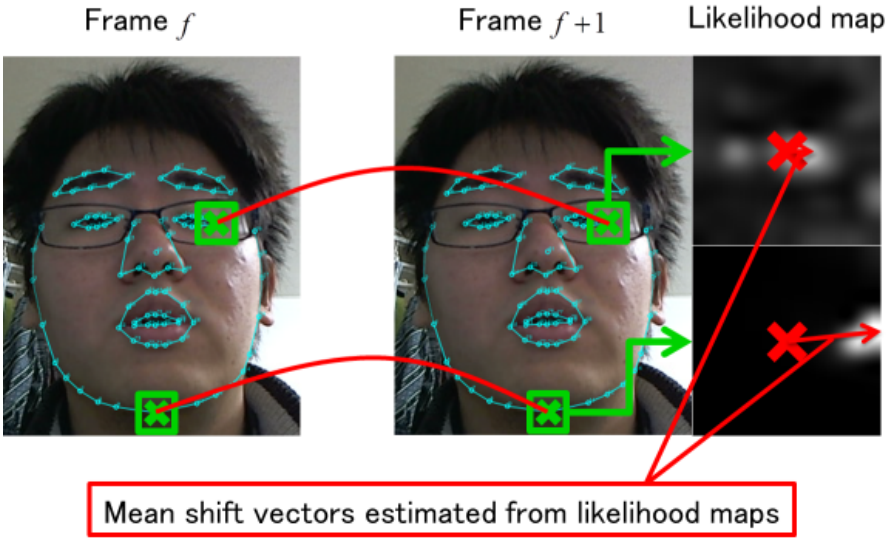

Fig. 2. The likelihood of alignment at a particular landmark location. Mean-shift vectors from each landmark are computed using the likelihood of alignment and the parameters $\mathbf{p}$ are updated.

(a) Side view (b) Upward view

(c) Lighting change (d) Partial occlusion

(e) Self shadowing (f) Motion blur

Fig. 3. Facial expressions captured in the strict environment

as do the conventional facial image databases such as AFW [11], LFPW [12], HELEN [13], and i-bug [14]. Furthermore, many blurred images are included because Elfoid is assumed to be used in the hand as a mobile phone. Facial images captured in strict environments are shown in Fig. 3. It is possible to find that various types of images, such as an upward view image and a blurred image which are not included in the conventional image database, are included in our database.

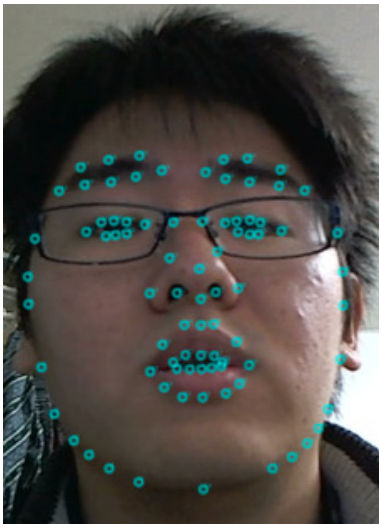

Fig. 4. Feature positions used in the experiment

4 Experiments

We have conducted accuracy verification experiments using various image features and classifiers to improve the face tracking method. The amount of movement of the feature point is estimated using the likelihood maps in Eq. (5). In Eq. (5) the classifier is generated using image features extracted at the manually annotated landmark locations. We conducted accuracy evaluations on our database, AFW [11], LFPW [12], HELEN [13], and i-bug [14]. This database, consisting of 500 images, is generated by extracting images from each database evenly and is annotated with an 83-point markup used as ground truth landmarks.

A total of 250 images, 50 from each database, were selected randomly and used as training data. Feature positions used in the experiment are shown in Fig. 4. As features, the gray scaled patch, image gradient, Local Binary Patterns (LBP) [5], Local Directional Pattern (LDP) [6], Local Phase Quantization (LPQ) [7], SIFT [8], SURF [9] and HOG [10] were used for producing the likelihood maps. Classifiers were generated using logistic regression and SVM. The size of the face region was normalized to 51x51 pixels. Image features were extracted at each feature position using 21x21 pixel regions. The size of the likelihood map was 11x11 pixels. The amount of movement of the feature point was estimated using the likelihood map generated from the results of the discrimination.

A total of 250 images that were not used as training data were used as test data. In the test phase, the Euclidean distance between the estimated position and the ground truth was used for an evaluation. As an evaluation of each facial part, Table 1 shows the number of optimal pairs of feature and classifier for discrimination. The results from Table 1 show that the optimal pairs of feature and classifier are different in each part of the face. To select the optimal feature and classifier by parts realizes accurate tracking.

Table 1. The number of optimal feature and classifier for discrimination in each part

feature	classifier	the number of features
Grayscale	Logistic regression	16
Grayscale	SVM	13
SURF	Logistic regression	12
SIFT	SVM	9
Gradient	Logistic regression	6
HOG	Logistic regression	6
SURF	SVM	6
HOG	SVM	4
LBP	SVM	3
LPQ	SVM	3
LDP	SVM	2
LBP	Logistic regression	1
LDP	Logistic regression	1
SIFT	Logistic regression	1
LPQ	Logistic regression	0
Gradient	SVM	0

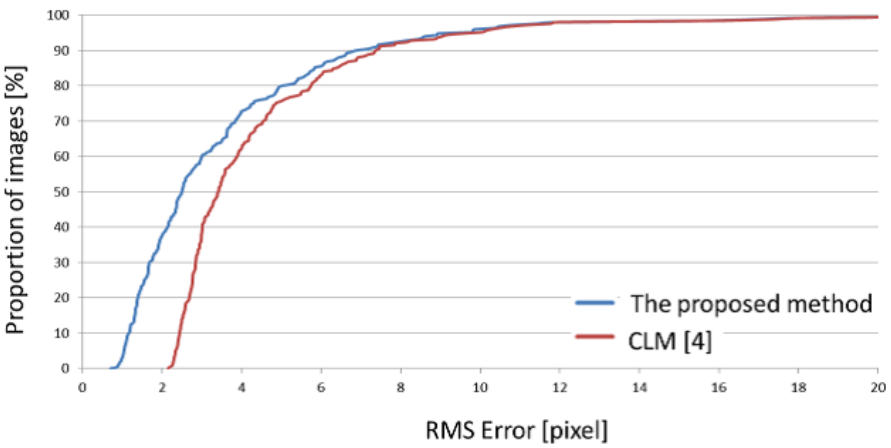

Fig. 5. Comparison between representative CLM [4] and the proposed method

We used the representative CLM [4] as the baseline method for comparison in this experiment. Figure 5 shows the comparison results. The proposed method used the combination of the features and classifiers in Table 1. The representative CLM [4] used image gradient and logistic regression. In Fig. 5, the horizontal

axis shows RMS error and the vertical axis shows proportion of images. It can be seen that the proposed method outperforms the representative CLM [4].

5 Conclusions

We present local feature evaluations for the CLM framework. We conducted accuracy verification experiments using various image features and classifiers. In our study, tracking accuracy can be improved by changing the features and classifiers for different parts of the face. In future work, we will implement a face tracking system that can switch the features and classifiers adaptively, responding to changes in the environment.

Acknowledgments. This research was supported by the JST CREST (Core Research for Evolutional Science and Technology) research promotion program "Studies on cellphone-type teleoperated androids transmitting human presence."

References

1. Becker-Asano, C., Ogawa, K., Nishio, S., Ishiguro, H.: Exploring the uncanny valley with Geminoid HI-1 in a real-world application. In: Int'l Conf. Interfaces and Human Computer Interaction, pp. 121–128 (2010)
2. Tsuruda, Y., Hori, M., Yoshimura, H., Iwai, Y.: Generation of facial expression emphasized with cartoon techniques using a cellular-phone-type teleoperated robot with a mobile projector. In: Kurosu, M. (ed.) HCII/HCI 2013, Part V. LNCS, vol. 8008, pp. 391–400. Springer, Heidelberg (2013)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, p. 484. Springer, Heidelberg (1998)
4. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int'l Journal of Computer Vision **91**, 200–215 (2011)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Analysis and Machine Intelligence **24**, 971–987 (2002)
6. Jabid, T., Kabir, M.H., Chae, O.: Robust facial expression recognition based on local directional pattern. ETRI Journal **32**(5) (2010)
7. Ahonen, T., Rahtu, E., Heikkila, J.: Recognition of blurred faces using local phase quantization. In: Int'l Conf. Pattern Recognition, pp. 1–4 (2008)
8. Lowe, D.G.: Object recognition from local scale-invariant features. Proc. Int'l Conf. Computer Vision **2**, 1150–1157 (1999)
9. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
10. Dalal, N.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 886–893 (2005)
11. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2879–2886 (2012)

12. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Analysis and Machine Intelligence **35**, 2930–2940 (2011)
13. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012)
14. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision (ICCV) Workshops (2013)

http://www.springer.com/978-3-319-13736-0

Face and Facial Expression Recognition from Real World Videos
International Workshop, Stockholm, Sweden, August 24, 2014, Revised Selected Papers
Ji, Q.; B. Moeslund, Th.; Hua, G.; Nasrollahi, K. (Eds.)
2015, X, 145 p. 55 illus., Softcover
ISBN: 978-3-319-13736-0