# Motion Field Histograms for Robust Modeling of Facial Expressions

Tanzeem Choudhury and Alex Pentland
MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139, USA
{tanzeem,sandy}@media.mit.edu

## Abstract

*This paper presents motion field histograms as a new way of extracting facial features and modeling expressions. Feature are based on local receptive field histograms, which are robust against errors in rotation, translation and scale changes during image alignment. Motion information is also incorporated into the histograms by using difference images instead of raw images. We take the principal components of these histograms of selected facial regions and use the top 20 eigenvectors for compact representation. The eigencoefficients are then used to model the temporal structure of different facial expressions from real-life data in the presence of translational and rotational errors that arise from head-tracking. The results demonstrate a 44% average performance increase over traditional optic flow method for expressions extracted from unconstrained interactions.*

## 1. Introduction

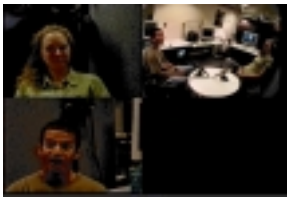
The aim of this research is to extract and process facial features from natural conversation sequences for the purpose of modeling emotional or cognitive states that generate different expressions. It is crucial that the system be capable of dealing with the unconstrained nature of real life data. We divide our task into four parts : (i) Data collection (ii) Head tracking and initial normalization (iii) Robust feature extraction (iv) Temporal modeling of multi-level expressions. The data collection process is designed to allow the natural flow of interactions. The system starts by performing initial normalization and alignment on the recorded data using a 3D model-based head tracker. However, the normalization and alignment is at best approximate and always suffers from errors in rotation, translation and scale. We have found no head-tracker that can provide sub-pixel accurate tracking for extended periods on medium-resolution video of natural, completely unconstrained head motion. Thus, it is important to select features that are robust against scale changes and failures of precise alignment of the input image, and which are stable and consistent over time. We were inspired by the performance of the local receptive field histograms for object recognition originally developed by Schiele and Crowley [12]. We extend the local histograms approach to be able to capture the fine scale changes in facial features and be suitable for building temporal models using Hidden Markov Models.

Most work in automatic understanding of facial expressions has focused on classification of the universal expressions defined by Ekman [7]. These expressions are sadness, anger, fear, disgust, surprise, happiness and contempt. Thus, the algorithms were tailored towards building models to recognize the universal expressions from static images or video sequences [4, 8, 14]. Recently, some work is being done towards recognition of individual action units that measure muscle action, proposed by Ekman as the basis for Facial Action Coding (FACS) [1, 5, 6]. All the experiments done and models built for facial actions or expressions require precise image registration and in some cases temporal alignment [6]. The image sequences used for these experiments depict very discrete and clean examples of specific action units or expressions which are almost impossible to find in natural, unconstrained interactions.

## 2. FACEFACTS: Modeling Natural Facial Expressions

Instead of simply assuming a representational basis developed for humans examining static imagery (e.g., FACS) our approach to expression modeling is to find a set of basis unit for expressions that can be extracted from natural data with high accuracy and then to use these basis expressions to build a representation of higher-level, complex expressions. In this section we describe all the elements of our proposed framework for analysis and modeling of facial expressions. We start with the design of our data collection environment. Then, we elaborate on the head-tracking, feature extraction

(a) A conversation scene



(b) Input to the Hi8 recorder

**Figure 1. Conversation Space Set-up**

and modeling methods that allow us to study and analyze naturally expressive data.

## 2.1. Data Collection

The data collection process should allow the subjects to move freely and naturally. The cameras and microphones should be placed in a way that is very unobtrusive, yet able to capture the changes in facial and vocal expressions. To fulfill this requirement, we designed and built a conversation space in our lab. The resources used to build the conversation space are : (i) Three Sony EVI-D30 pan-tilt cameras (ii) Two AKG 1000 cardioid microphones (ii) Quad splitter (iv) Hi8 video recorder (v) TV monitor (vi) Table and a set of chairs (vii) Cables, connectors, camera mounts etc.

Two of the cameras were directly behind and slightly above each subject to get a frontal view of the subjects participating in the conversation. To capture the entire scene, a wide angle lens was attached to the third camera and placed in the center corner. The outputs of the cameras were fed into a quad splitter. The quad splitter was connected to the recorder and the TV screen was used to monitor the recording process and to detect any problems. The whole system is almost invisible to the subjects and does not constrain them in any way.

## 2.2 Head Tracking and Feature Extraction

In order to analyze unconstrained video which includes significant head movement, it is necessary to know changes in the head pose as well as the facial features. We used three existing 3D model-based head-trackers developed by our group [2, 10, 13]. We use the output to normalize and warp the face to a frontal position. In our experience no existing head tracker is able to track unconstrained data for a extended period of time, so the output normalized images have errors in position and scale. Consequently, it is very important to build in robustness into the extracted features. Evidence suggests that when people are engaged in a conversation the most frequently occurring movements are raising the eye-brow, lowering the eye-brow and some form of smiling [9]. Thus, we decided to automatically extract the eyebrows, eyes and mouth region from the normalized images.

### 2.2.1 Motion Field Histograms

Lack of accuracy in normalization of face images during head-tracking leads to difficulty in recognizing changes in facial features. Therefore, features that are less sensitive to small position and scale changes are likely to prove more reliable for our task.

In choosing our features, we were inspired by the object recognition system proposed by Schiele and Crowley [12]. Objects are represented as multidimensional histograms of vector responses of local operators. Schiele experimentally compared the invariant properties of a few receptive field functions, including Gabor filter and local derivative operators. His results showed that Gaussian derivatives provided the most robust and equivariant recognition results. Given the Gaussian distribution $G(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$, the first derivative in the $x$ and $y$ is: $Dx(x,y) = -\frac{x}{\sigma^2}G(x,y)$ and $Dy(x,y) = -\frac{y}{\sigma^2}G(x,y)$. The Laplace operator is $Lap(x,y) = G_{xx}(x,y) + G_{yy}(x,y)$, where $G_{xx}(x,y) = (\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2})G(x,y)$, $G_{yy}(x,y) = (\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2})G(x,y)$.

In our experiment, we used only the first derivative and the Laplacian at two different scales resulting in a 6 dimensional histogram. The resolution of the histogram axis was either 16 or 32 pixels. For more details on creating the histograms please refer to [12].

The probability of an object $O_n$ given local measurement $M_k$ is obtained using Bayes' rule:

$$p(O_n|M_k) = \frac{p(M_k|O_n)p(O_n)}{p(M_k)}$$

where $p(O_n)$ is the prior probability of the object which is known and $p(M_k)$ as the prior probability of the filter output which is measured as $\sum_i p(M_k|O_i)p(O_i)$. So, $p(M_k|O_i)$, the probability density of an object $O_n$ differs from the multi-dimensional histogram of an object by a normalization term. If we have K independent measurements $M_1, M_2, ....., M_K$ then the probability of the object $O_n$ is:

$$p(O_n|M_1, M_2, ..., M_K) = \frac{\prod_k p(M_k|O_n)p(O_n)}{\prod_k p(M_k)}$$

To ensure independence between measurements, we choose the minimum distance $d(M_{k_1}, M_{k_2}) \geq 2\sigma$ between two measurements $M_{k_1}$ and $M_{k_2}$. The measurement locations can be chosen arbitrarily and it is not necessary have measurements at corresponding points and only a certain number of local receptive field vectors need to be calculated, the method is fast and robust to partial occlusion.

**Figure 2. Example blink sequence**

**Figure 3. Difference images of the sequence**

Because we are trying to distinguish changes in the same object as opposed to different objects we incorporate some motion cues into our histograms by using difference images, which significantly improve the performance. To capture both fast and slow changes, temporal differencing should be done at different rates. However, for short time scale or fast expressions it is enough to have consecutive frame differencing of images recorded at 30 frames/second. In the original framework of Schiele and Crowley [12] the histograms were compared directly using the $\chi^2$ statistic, histogram intersection, or mahalanobis distance. In our case, it is important to see changes in the histograms over time for an expression rather than compare histograms. Thus, to compactly represent the histograms and reduce the dimensionality for temporal modeling we take the PCA of the input histograms and use the top 20 eigenvectors, which capture 90% of the variance to represent the histogram space.

#### 2.2.2 Optic Flow

To compare the performance of the local histograms with existing methods used for expression modeling, we calculated the optic-flow of selected regions. The flow estimate is obtained with a multi-scale coarse-to-fine algorithm based on a gradient approach described by [3, 11]. Optic-flow is sensitive to the motion of image regions and the direction in which different facial features move, but it is also sensitive to the positions of the feature points.

We computed the dense optic flow of the images extracted from the normalized face image. Then we calculated the eigenvectors of the covariance matrix of the flow images. The top 20 eigenvectors were used to represent the motion observed in the image regions. These captured 85% of the variance. The eigencoefficients computed for each image region were the new input features. However, large tracking errors leading to improper normalization can cause optic-flow to provide misleading results as shown in Table 1.

### 2.3 Feature Modeling

After the feature extraction stage comes the modeling stage — in this stage we would like to capture the relationships among feature movements over time in expressive gestures. For example, an eye-blink always consists of an open eye gradually closing and then opening again — there is an explicit temporal pattern in the eyelid motion. There are many such cases where there is a clear temporal pattern, e.g during the raising of eyebrows or looking towards some direction etc. We need to capture these expressive patterns to build models for a person's facial expressions. However, a brow raise or an eye-blink does not provide enough information about emotional expressions — it is the combinations and the temporal relationships between the short time expressions that can explain what a person is trying to convey through his/her facial expressions.

Our low level modeling step uses Hidden Markov Models (HMMs) to model expressions that occur consistently across people and which can be extracted from unconstrained data. For eyes, these expressions are blinks, raising and lowering of eyebrows, looking in different directions etc. Once these low-level expressions can be detected reliably we can model high level expressions which can be described as a structured combination of the low-level expressions.

## 3. Results

The results presented in this section demonstrate that expression models based on local histograms outperform optic flow when the images are not perfectly aligned and normalized. However, they are at least as good as optic-flow based models for perfectly aligned images. If we are to build a system that can reliably extract information from natural interactions it has to be robust against tracking and normalization errors. The lowest level or the shortest time scale expressions that we model should be such that they can be extracted with a high degree of confidence from unconstrained video. These low level models can then be used to build mid level and high level structures that capture the emotional or cognitive meaning of a conversation or interaction.

Table 1 shows our results using local histograms and optic-flow for images obtained without compensating for head-tracker errors. In this case, we have both translational (between 5% - 10%) and rotational error ($\pm$5degrees) and minor scale changes. We used 44 expression sequences in total for training and 38 for testing consisting of 1719 and 1417 difference images respectively depicting the following eye expressions: blink, looking left, looking right, raising the eyebrow and looking up. Images were recorded at 30 frames/second, the mean length of a typical expression was 39 frames with a standard deviation of 19 frames. Expres-

| Expression | Local Histograms | Optic - Flow |
|---|---|---|
| Brow Raise | 90.0 % | 88.9 % |
| Blink | 90.0 % | 40.0 % |
| Right | 90.0 % | 25.0 % |
| Left | 100.0 % | 20.0 % |
| Up | 100.0 % | 71.4 % |

**Table 1. Recognition rates — images with tracking errors**

| Expression | Local Histograms | Optic - Flow |
|---|---|---|
| Brow Raise | 90 % | 100 % |
| Blink | 90 % | 90 % |
| Right | 100 % | 100 % |
| Left | 100 % | 100 % |
| Up | 100 % | 100 % |

**Table 2. Recognition rates — perfectly normalized images**

sions were not normalized to a constant length, unlike in the study by Donato et al. [6], as the variation in length contains important information as well. For example, a long blink might happen because a person is drowsy as opposed to regular short blinks. Training and testing expressions were obtained from two separate data recordings on two different days. All the expressions were trained on the histogram coefficients or the flow coefficients using three-state left-to-right HMMs.

Table 2. shows our results using local histograms and optic-flow for images that were accurately normalized. Again, all the expressions were trained using three-state left-to-right HMMs. On average we had 10 sequences/per expression for both training and testing.

## 4. Conclusion

One of major stumbling blocks in expression modeling from unconstrained data has been inaccurate alignment and normalization of face images. Thus, models were built using constrained and unnatural data which is not scalable to real life cases. In this paper we proposed an expression modeling technique that can work robustly with natural data and have high recognition accuracy. Our models are built on feature based on motion motion field histograms, which have robustness against errors in rotation, translation and scale changes during image alignment. The results demonstrate a 44% average performance increase over traditional optic flow methods for expressions extracted from unconstrained interactions.

## 5. Acknowledgments

The authors are very grateful to Bernt Schiele for all his help and for the use of his code. This work builds on the receptive field histogram based object recognition system introduced by him. We also thank Sumit Base, Tony Jebara and Jacob Strom, who developed the head trackers used in our experiments.

## References

[1] M. Bartlett, J., P. Ekman, and S. T. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263, March 1999.

[2] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head-tracking. In *13th International conference on Pattern Recognition*, Vienna, Austria, August 25-30 1996.

[3] J. Bergen, P. Anandan, and H. R. Hierarchial model-based motion estimation. In *Second European Conference on Computer Vision*, pages 237–252, Santa Margherita, Italy, May 1992.

[4] M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. Journal on Computer Vision*, 25(1):23–48, 1997.

[5] J. Cohn, A. Zlochower, J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Journal of Psychophysiology*, 36(1):35–43, January 1999.

[6] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[7] P. Ekman and W. Friesen. *Unmasking the Face*. Prentice-Hall, 1975.

[8] I. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. In *CVPR*, pages 76–83, 1994.

[9] J. Hager and P. Ekman. Essential behavioral science of the face and gesture that computer scientists need to know. In *International Workshop on Automatic Face and Gesture Recognition*, June 1996.

[10] T. Jebara and A. Pentland. Parameterized structure from motion for 3D adaptive feedback tracking of faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[11] B. Lucas. and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Image Understanding Workshop*, pages 121–130, 1981.

[12] B. Schiele and J. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *Proceedings of International Conference on Pattern Recognition*, Vienna, Austria, August 1996.

[13] J. Strom, T. Jebara, S. Basu, and A. Pentland. Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. In *International Conference on Computer Vision: Workshop on Modelling People*, Corfu, Greece, September 1999.

[14] Y. Yacoob and L. Davis. Computing spatio-temporal representation of human faces. In *CVPR*, Seattle, WA, June 1994.