Representation and Recognition of Complex Human Motion

Jesse Hoey James J. Little
Department of Computer Science
University of British Columbia
Vancouver, British Columbia V6T 1Z4
jhoey@cs.ubc.ca little@cs.ubc.ca

Abstract

The quest for a vision system capable of representing and recognizing arbitrary motions benefits from a low dimensional, non-specific representation of flow fields, to be used in high level classification tasks. We present Zernike polynomials as an ideal candidate for such a representation. The basis of Zernike polynomials is complete and orthogonal, and can be used for describing many types of motion at many scales. Starting from image sequences, locally smooth image velocities are derived using a robust estimation procedure, from which are computed compact representations of the flow using the Zernike basis. Continuous density hidden Markov models are trained using the temporal sequences of vectors thus obtained, and are used for subsequent classification. We present results of our method applied to image sequences of facial expressions both with and without significant rigid head motion and to sequences of lip motion from a known database. We demonstrate that the Zernike representation yields results competitive with those obtained using principal components, while not committing to specific types of motion. It is therefore ideal as a fundamental building block for a vision system capable of classifying arbitrary motion types.

1 Introduction

This paper describes a representation of flow in image sequences of complex human motion. The main contribution of this work is to demonstrate the basis of Zernike polynomials as a representation of complex flow fields which is both simple and general. It can be applied to any type of motion at any scale, and so can be used with little or no prior knowledge about the content of image sequences. The low order (0 and 1) Zernike polynomials correspond to the standard affine basis. The next order polynomials correspond to extensions of the affine basis, roughly *yaw*, *pitch* and *roll*, as explored in [3, 12]. Higher orders represent motions with higher spatial frequencies. The basis is orthogonal over the unit disk, such that each order can be used as an independent characterization of the flow, and each flow field has a unique decomposition in the Zernike basis.

The human visual system is remarkable in its ability to accurately recognize a wide variety of motion types. In particular, the recognition of human action is important from an evolutionary standpoint since it gives information about social interactions. However, human actions come in many shapes and span many scales of motion from walking gaits to the fine scale motions of individual muscles in the face. A central question is therefore how the visual system can accurately represent such different types of motion. The accomplishment of this task depends on the tradeoff between segmentation and representation. That is, one or multiple regions of interest and scales must be identified in the visual field and the scales of these regions determine the complexity of the representation to be used, larger scales requiring higher order features. We propose a representation of flow which is scale invariant and hence can accomplish this task more readily than other commonly used approaches. The human face provides motion fields which can be analyzed at many scales, and we focus our effort on the understanding of human facial motion in order to test our approach.

Related previous work in facial expression recognition can be divided into holistic approaches [14, 15, 12] and local (segmented) approaches [3, 6, 17, 19, 11] .

Holistic approaches look at the entire face, as we do, but develop specific basis sets for each type of motion. Turk and Pentland [14] used principal components analysis (PCA) to find a set of basis eigenvectors describing variation across static face images. Belhumeur *et al.* [15] compared and contrasted PCA with Fisher linear discriminants in a similar approach. While these methods are not specifically developed for a certain type of motion, they generate classifiers which are. That is, the generated basis sets are specific to a particular recognition task, and further, to a particular scale. Scale invariance is achieved only through image warping or storing basis sets over a range of scales. Morimoto *et al.* [12]

used optical flow over the entire head modeled with a low order polynomial basis in an approach similar to the one proposed here. Their system was designed to recognized large rigid head motions and used a rule-based method specific to head motion for classification.

Segmentation methods can be used to break the face images into component parts such as eyes, nose and mouth. Segmentation facilitates the modeling of motion, since within each region the flow can be more accurately represented using a low order basis. At one extreme are methods which use models of the facial muscles [6]. A variety of less constrained flexible models have been studied in the context of face recognition [9, 17, 19, 11]. At the other extreme, Yacoob and Davis [18] use very little modeling, segmenting the face and developing statistical descriptions of the changes of the segmented regions. Black and Yacoob [3] used mid-level representations of flow over small image regions. This work is related to ours, in that they model flow using the lowest order Zernike polynomials.

Segmentation methods rely on some geometric information about head shape and motion, and then use low spatial frequency models of motion and structure across small regions. PCA-based approaches, on the other hand, use a range of spatial frequency templates which span large areas. Our method fits snugly in between these two by using universal templates of flow over a range of spatial frequencies and at any spatial scale. It is distinguished by its lack of dependence on *a-priori* information about the type of motion being observed. It can be applied, as in this paper, to an entire motion region, thus imitating a PCA-based approach, without the computational burden of generating a motion-specific basis. It can also be applied at smaller scales, thus reducing to the segmented approaches [3]. Finally, the Zernike representation could be used in a multiscale approach, the vectors from each level driving segmentation at the next level, thus completely bridging the gap.

Zernike polynomials, originally developed for the modeling of lens aberrations, have been used in recognition tasks such as handwriting, aircraft outlines [1], and hand poses [7]. To our knowledge, Zernike polynomials have not been applied to optical flow fields. However, other types of polynomials, such as the central moments [10] have been.

Our method starts with an estimation of the flow fields for each subsequent pair of images in a sequence using the robust gradient-based regularization method of [2]. The centroid and scale of the area of interest is then estimated, and a feature vector is obtained by projecting the flow onto the basis of Zernike polynomials. Over the course of the sequence, this feature vector describes some trajectory through the Zernike vector space. Continuous density hidden Markov models are trained on the temporal sequences of these Zernike vectors, and are used for subsequent classification tasks. Three datasets are examined: facial expressions without rigid head motion, facial expressions with rigid head motion, and lip-reading sequences from the *Tulips1* database [13]. Our results from the second dataset are compared with results obtained using a basis of principal components. Despite expectations to the contrary, we find our representation can outperform the principal components one, indicating that a universal basis set, from which a more general motion classifier can be built, is feasible.

This paper will proceed as follows. Section 2 presents the robust flow method and the Zernike polynomial basis. Section 3 reviews hidden Markov models. Our experiments are presented in Section 4, followed by a discussion of the results and of future plans in Section 5.

2 Data

Data were gathered using a Sony pan-tilt EVI-D30 camera and Matrox Meteor frame grabber which recorded sequences of 160x120 greyscale images taken at 30Hz. The following describes the method of deriving the ZP feature vectors from these image sequences.

2.1 Flow fields

Flow fields were generated using the robust gradient-based regularization method of [2]. The robust method involves parameters controlling the resulting smoothness of the flow. As our method can represent high spatial frequencies, the amount of smoothing plays an important role, which has not been fully investigated yet, but is partially addressed in section 4.3.

2.2 Zernike polynomials

Zernike polynomials are an orthogonal set of complex polynomials defined on the unit disk. They arise in the expansion of a wavefront function for optical systems with circular pupils [8]. They are expressed in polar coordinates ($\rho = \sqrt{x^2+y^2}$, $\phi = \arctan(y/x)$) as a radial function, $R_n^m(\rho)$, modulated by a complex exponential in the angle, ϕ, as follows:

$$U_n^m(\rho,\phi) = R_n^m(\rho)e^{im\phi} \tag{1}$$

with radial function, $R_n^m(\rho)$, given by

$$R_n^m(\rho) = \sum_{l=0}^{(n-|m|)/2} \frac{(-1)^l(n-l)!}{l![\frac{1}{2}(n+|m|)-l]![\frac{1}{2}(n-|m|)-l]!}\rho^{n-2l}$$

for n and m integers with $n \geq |m| \geq 0$ and $n-m$ even. The first few radial basis functions are therefore:

$$R_0^0 = 1 \qquad R_1^1 = \rho \qquad R_2^0 = \rho^2$$
$$R_2^2 = 2\rho^2 - 1 \qquad R_3^1 = 3\rho^3 - 2\rho \qquad R_3^3 = \rho^3$$

The Zernike polynomials are orthogonal on the unit disk, and obey the following orthogonality relation:

$$\int_0^1 \int_0^{2\pi} U_n^m(\rho,\phi) U_{n'}^{m'}(\rho,\phi)\rho\, d\phi\, d\rho = \frac{\pi}{(n+1)}\delta_{nn'}\delta_{mm'}, \quad (2)$$

where $\delta_{nn'} = 1$ if $n = n'$, and 0 otherwise. We wish to exploit this orthogonality in recognition tasks. It allows the decomposition of an arbitrary function on the unit disk, $F(\rho,\phi)$, in terms of a unique combination of odd and even Zernike polynomials. That is, [8]

$$F(\rho,\phi) \approx \sum_{m=0}^{M}\sum_{n=m}^{N} [A_n^m cos(m\phi) + B_n^m sin(m\phi)]\, R_n^m(\rho), \quad (3)$$

which can be used to approximate a sufficiently smooth function $F(\rho,\phi)$ to any degree of accuracy by making N and M large enough. Using the orthogonality relation (Equ. 2), the coefficients A_m^n and B_m^n can be obtained as

$$\begin{matrix} A_n^m \\ B_n^m \end{matrix} = \frac{\epsilon_m(n+1)}{\pi}\int_0^1\int_0^{2\pi} F(\rho,\phi) R_n^m(\rho) \begin{matrix} cos(m\phi) \\ sin(m\phi) \end{matrix} \rho\, d\phi\, d\rho, \quad (4)$$

where

$$\epsilon_m \equiv \begin{cases} 1 & \text{if } m = 0 \\ 2 & \text{otherwise} \end{cases}$$

We chose to use the Zernike polynomials because of their orthogonality properties, which leads to greater ease of interpretation of the resulting feature vectors. Other possibilities, such as the generating functions of the central moments, x^p and y^q, do not have these properties.

2.3 Flow representation using ZPs

Zernike polynomials are defined on a disk, and so a circular area within each flow image must be identified which will be projected onto the Zernike basis. This is accomplished by manually specifying a scale and centroid for the first frame of each sequence, and then using the first order (affine) components of the flow fields to track the region. Centroids are specified for face images by taking the center of the face from hairline to chin and between the outermost visible edges of the face. Scales for face images are measured along the horizontal direction. Centroids for the lip sequences are taken in the approximate midpoint of the mouth, with scale as width of the image.

Once a scale and centroid have been identified for each flow image, the horizontal and vertical velocities $u(x,y)$, $v(x,y)$ are projected onto the Zernike basis using the discrete equivalent of Equation 4:

$$\begin{matrix} {}^u A_n^m \\ {}^u B_n^m \end{matrix} = \frac{\epsilon_m(n+1)}{\pi}\sum_x\sum_y u(x,y) R_n^m(\rho) \begin{matrix} cos(m\phi) \\ sin(m\phi) \end{matrix} \quad (5)$$

where $\rho = \sqrt{x^2+y^2} \le c$, c is the scale factor and x, y are coordinates relative to the centroid, and $\phi = \arctan y/x$. The flows can be reconstructed from the coefficients using Equation 3. The region tracking update equations are defined using the first order coefficients:

$$x_c' = x_c + {}^u A_0^0 \quad y_c' = y_c + {}^v A_0^0 \quad r' = r + \frac{1}{2}({}^u A_1^1 + {}^v B_1^1)$$

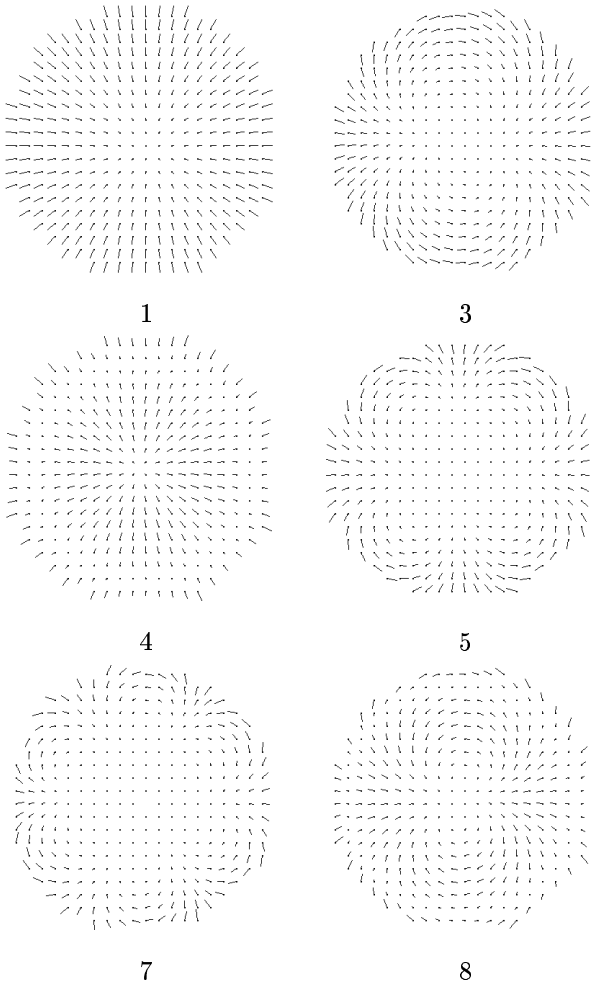
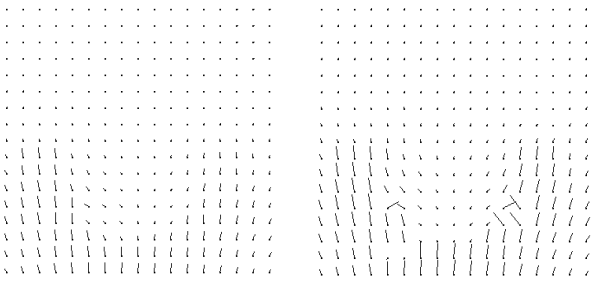


Figure 1. Example flows generated by the $\kappa = 1,3,4,5,7,8$ ZPs.

In this study, we are interested in the structure of the motion, and not the true value. That is, a rapidly expanding smile will be considered the same as a slowly expanding one. Therefore, we normalize all feature vectors to unit length. Feature vectors are sets of normalized $\{{}^u A_n^m, {}^u B_n^m, {}^v A_n^m, {}^v B_n^m\}$ coefficients, with n, m assuming a numeric ordering obeying $n - m$ even and $n \ge m$: $\{(0,0),(1,1),(2,0),(2,2),(3,1),(3,3),...\}$. The four scalars making up each (n,m) combination will be called ZPs in what follows, and will be referred to by the position indices, $\kappa = \{0,1,...,N\}$, of this ordering. Basis sets will be denoted $\{N_0,...,N\}$, or simply by N, in which case the basis set is $\{0,...,N\}$. Although the angular spatial frequency (m) is hidden by this indexing, higher values of κ indicate higher radial spatial frequencies. Figure 1 shows examples of flows generated by the first few ZPs.

0 1 2

Figure 2. First three frames from the beginning of a *happy* expression sequence.

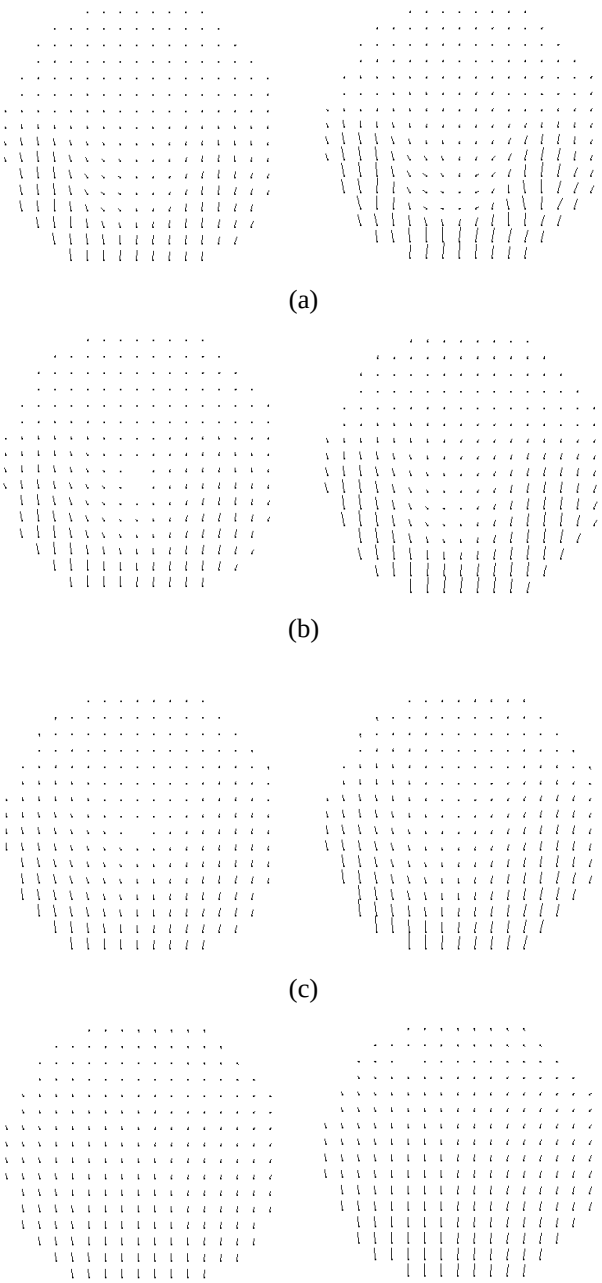


$0-1$ $1-2$

Figure 3. Flow fields derived from the sequence in Figure 2, for image pairs $0-1$ (left) and $1-2$ (right).

2.4 Accuracy of the representation

The Zernike basis, being complete and orthogonal, can provide an exact representation of any flow field projected onto it. In practice, a small subset of ZPs must be chosen which can accurately represent the types of flows being represented, as described in the last section. In order to show how the ZPs represent flows, consider the image sequence shown in Figure 2. This sequence shows the beginning of a *happy* expression, characterized by a horizontal expansion and vertical raising of the corners of the mouth. The flow fields for this sequence are shown in Figure 3. The ZP decomposition is performed on the above flow fields using Equation 5, and the flows are reconstructed using Equation 3, with $N = 1$, $N = 6$, $N = 13$ and $N = 48$, as shown in Figure 4. Comparison of Figures 3 and 4 shows the effects of adding higher order ZPs to the reconstruction. With $N = 48$ the reconstruction is nearly perfect. Note that the spatial frequency of the Zernike basis at $\kappa = 48$ ($n = m = 12$) is high enough to partially capture the localized aberrations in the flow field. The reconstruction with $N = 13$ is very close to the higher order one, showing that the flows in these images do not in fact have very high spatial frequency components, except for the aberrations in the flow, which are no longer captured by the $N = 13$ basis elements. With $N = 7$, the reconstructed flow is degraded from the original, with distortions happening primarily just below the the center of the disc. With $N = 1$ (affine basis), the reconstruction is poor, although the spreading of the mouth towards the bottom of the image can still be observed, as can the lack of flow in the top part of the image. These cues will be shown to be adequate in classification



(a) (b) (c) (d)

Figure 4. Flow fields from Figure 3 reconstructed using $N = 48$ (a), $N = 13$ (b), $N = 6$ (c), and $N = 1$ (affine) (d).

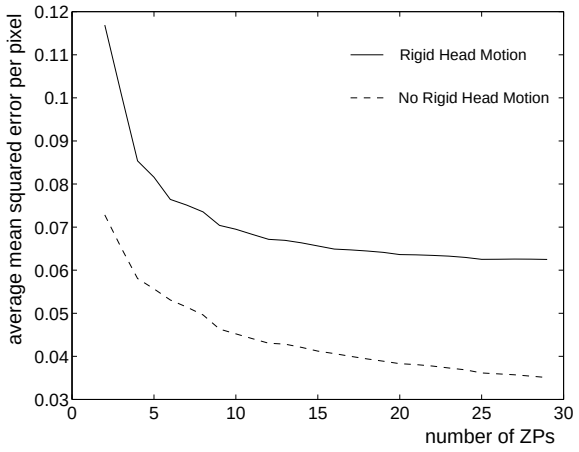


Figure 5. Mean squared error per pixel plotted as a function of the number of ZPs used for flow reconstruction, for the datasets of image sequences with little rigid head motion (dashed line) and with significant rigid head motion (solid line).

tasks where the facial expressions to be classified are sufficiently distinct.

To obtain a quantitative measure for the distortion as a function of the number of Zernike Polynomials used in the basis, we compute the mean squared error per pixel between the original flows and the flows reconstructed from the Zernike coefficients, and plot the error averaged over all images as a function of the number of ZPs used in the reconstruction. Figure 5 shows this for two of the datasets of facial expressions considered in Section 4. The images with rigid head motion (solid line) show a reconstruction error which drops off sharply with the addition of the first 10 orders, but little decrease is observed beyond $N = 15$, in agreement with the qualitative analysis presented in Figures 3 and 4. The flows generated in the second dataset are smaller in magnitude due to the lack of rigid head motion, but we see the distortion again drops off with the number of ZPs used in the reconstruction.

3 Temporal modeling

A motion sequence produces a feature vector which describes some trajectory through the Zernike vector space. It is our premise that the motion is produced by some underlying hidden process, which can be modeled as a sequence of states, each of which is responsible for a class of feature vectors. That is, the flows are roughly constant over small intervals of time. Therefore, we use a hidden Markov model (HMM) with continuous observation probabilities in each state modeled with a unimodal Gaussian distribution. Models are built from training data using the standard expectation-maximization procedure [16]. The number of hidden states is an important parameter for the performance of HMMs, which we investigate in Section 4. To classify a new sequence of observations, $\mathcal{O}$, we calculate the probability of the observation sequence given each model, $\mathcal{M}$. Then, the model with the highest likelihood is the one which the new sequence of observations is assigned.

4 Experiments

Three experiments were carried out, the first two involving a single subject performing five facial expressions *disgust, fear, happiness, sadness* and *surprise*. The expressions were performed from beginning through apex to ending. The last experiment was done on the Tulips1 database [13], which involves 12 subjects saying 4 words twice each.

The first experiment involved taking image sequences of a test subject performing one of the five facial expressions multiple times in succession, while keeping rigid head motion to a minimum. For each expression, 10-15 sequences of 20-30 frames in length were recorded. The derived ZPs from all sequences but one were used to train models, and the left-out sequences were classified, known as a *leave-one-out* validation method.

The second experiment also consisted of image sequences of a single subject performing the five expressions, but without any restrictions placed on head motions. In fact, a conscious effort was made by the test subject during this experiment to make head motions as uniform across expression sequences as possible. Rigid head motions in this experiment included translations, looming and retreating, yaw, pitch, and roll. For each expression, 10-15 sequences of 30-40 frames in length were recorded. The *leave-one-out* procedure was again carried out on the derived ZPs from these sequences.

In a third experiment, we used the Tulips1 database [13] of lip motion sequences. These sequences were taken from 12 subjects (9 male, 3 female) each saying the first four digits in English twice. The sequences are much shorter than our first two experiments, some having only 3 images.

4.1 Small head motions

Figure 6 shows the first 3 frames of a *happy* expression, with little rigid head motion. Percentage results of the *leave-one-out* procedure are shown in Table 1. We present results for various combinations of ZPs and using HMMs of different sizes, in order to characterize the effects of these parameters. Table 1 shows the basis sets split into three groups vertically, ordered according to the lowest order polynomial in the set. The cardinality of the basis set

0 1 2

Figure 6. First three frames from the beginning of a *happy* expression sequence, with little rigid head motion.

determines the ordering the bases within each group. The first group of bases include the translational components alone, ($\{0\}$); the affine basis, ($\{01\}$), which characterizes planar flows; the extended affine basis, ($\{0-4\}$), which describes perspective distortions such as *pitch*, *roll* and *yaw*; and a further extension ($\{0-7\}$), which introduces higher spatial frequencies (up to angular frequency $m = 3$). We trained left-right HMMs with 1,2,4 and 6 state hidden variables. HMMs with 1 hidden state are not really HMMs, but simply compare the sequences by temporally averaging the ZPs over the entire sequence, and then comparing using a covariance-weighted distance metric, or Mahalanobis distance. In contrast to most approaches, we decided to examine the performance of our method across different sizes of HMMs. This provides an interesting tradeoff between the complexity of the representation and the complexity of the temporal model.

The results in Table 1 show that the five expressions of a single subject are well modeled by using an affine description of the flow across the entire face. This is to be expected, since the expressions used have simple characterizations in terms of the entire face. For example, *happiness* consists of a general raising of the features (a simple translation), whereas *disgust* is a general contraction (an affine flow), and surprise an expansion (also affine). The higher order basis sets also give slightly better performance rates, but, with little room for improvement, the distinction is not great. The hidden Markov models do not provide any clear advantage in this first experiment, and can even reduce the recognition rates in some cases. Consider the ($\{0\}$) basis, in which recognition drops from 75% with one state to 66% with two. With only a single state, the mis-classifications are evenly spread over the five expressions. We found that adding a second state serves to separate the *happy*, *sad* and *surprise* expressions, but increases confusion between *fear* and *disgust*. The overall temporal averages of these two expressions are more distinguishable than when broken into two parts. The reduction in recognition rates for the higher-dimensional basis sets when more hidden states are added is due to the increased complexity of the model leading to a more singular training and classification procedure.

	hidden variable states			
basis	1	2	4	6
$\{0\}$	74	66	72	79
$\{0\ 1\}$	98	98	95	97
$\{0\text{-}4\}$	100	100	98	98
$\{0\text{-}7\}$	100	100	98	98
$\{2\}$	74	62	81	84
$\{2\ 3\}$	95	97	95	98
$\{2\text{-}5\}$	100	100	100	100
$\{2\text{-}9\}$	100	100	100	97
$\{4\}$	88	93	89	92
$\{4\ 5\}$	94	92	91	91
$\{4\text{-}7\}$	100	98	97	95
$\{4\text{-}11\}$	100	100	98	89

Table 1. Test sequence percentage results on 5 facial expressions *disgust, fear, happy, sad, surprise*, **on image sequences with minimal rigid head motion involving a single test subject.**

Figure 7. Three frames from *sad* **(top row), and from** *surprise* **(bottom row) expressions sequences involving rigid head motion.**

4.2 Rigid head motions

Figure 4.2 shows 3 frames of a *sad* and of a *surprise* expression, with significant rigid head motion. The rigid motions include translation, expansion, contraction, looming, and yaw, and are captured by the first four orders of the Zernike basis [3], $\kappa = 0 - 3$, and so basis sets such as the affine basis will have more difficulty in classifying facial expressions when combined with arbitrary rigid head motion. Table 2 shows the results using a variety of different combinations of ZPs. The higher order polynomials, in conjunction with the lower orders, do provide a distinct advantage in this case, as the ($\{0-7\}$) basis yields a 95% rate, while the affine basis only yields 74%. Single ZPs perform poorly, although the higher order ones do give slightly better results. Recognition with *only* higher orders does not perform as well as when the low orders are included. This emphasises the importance of the $\kappa = 0 - 3$ polynomials

	hidden variable states			
basis	1	2	4	6
{0}	27	40	35	33
{0 1}	66	69	64	74
{0-4}	86	86	89	87
{0-7}	95	94	94	87
{ 2 }	30	49	43	37
{2 3}	69	66	78	74
{2-5}	88	91	89	86
{2-9}	94	92	89	81
{ 4 }	61	58	55	60
{4-5}	80	72	79	80
{4-7}	92	92	94	92
{4-11}	88	89	87	84

Table 2. Test sequence percentage results on 5 facial expressions *disgust, fear, happy, sad, surprise* **, on image sequences with rigid head motion involving a single test subject.**

	hidden variable states			
basis	1	2	4	6
{ 2 }	35	37	39	32
{2 3}	75	85	85	80
{2-5}	85	89	88	88
{2-9}	91	91	83	75

Table 3. Results using PCA on the rigid head motion sequences. Compare to middle section of Table 2

	hidden variable states			
basis	1	2	3	
			rough	smooth
{0 1}	58	60	66	77
{0-4}	61	59	70	70
{0-7}	70	70	76	76
{2,4,8,9,10,14,22}	77	78	79	63

Table 4. Test sequence percentage results on the Tulips1 database, training on 11 of 12 speakers and testing on the twelfth. The smoothness factor of the flow regularization is indicated by *rough* **and** *smooth***.**

to the recognition task. These low orders have a great deal of expressive power in terms of flow fields, as is clear from their widespread use in pattern recognition. This experiment demonstrates that the extended versions of this basis are also useful for modeling flow.

The effects of the HMM size on the recognition rates are unclear in this experiment as well, primarily due to lack of training data. With $\kappa = \{0 - 7\}$, the observation vectors are modeled with a 26-dimensional Gaussian. Clearly, with as few as 300 observation vectors, splitting the space into six temporal regions is not advantageous. Lower dimensional feature vectors are more amenable to larger HMMs, but it appears that adding more than 2 hidden states does not yield any performance gain, probably due to overfitting of the data. A solution to these problems would be to extend the training step to include priors on model structure, which can help with both the overfitting problem and with the lack of training data. Minimum entropy methods, as explored in [4], are clearly a good starting point.

Principal components were also generated for this second dataset by scaling every second flow field to a single size (as required by the analysis), and computing eigenvectors ranked by the magnitude of their corresponding eigenvalues. These eigenvectors were then used as a basis on which flows were projected, the results of which were used to train HMMs, as before. Table 3 shows recognition results for using a subset of the principal components, starting from the third, and can be directly compared to the results in the middle section of Table 2. Similar results were obtained in comparison to the other two sections of Table 2. The PCA-based representation performs nearly as well in the best case, with improvements for the lower dimensional cases. These improvements arise because the lower order principal components have higher spatial frequencies than the low order Zernike polynomials, and can thus capture more complex differences. These results show that a well structured non-specific basis set can be used in place of the principal components, with competitive results.

4.3 Lipreading

The Tulips1 database [13] consists of each of 12 subjects saying the first four digits of English twice. In this case we use a *leave-one-speaker-out* procedure, where the models are trained using the data from 11 subjects, and remaining subject's eight test cases are classified using these models. Results are presented in Table 4 for 1,2 and 3 state HMMs, and four basis sets. The last basis set ($\{2, 4, 8, 9, 10, 14, 22\}$) was obtained by choosing the set of basis vectors which maximized the interclass distance, while minimizing the intraclass distance over the training images, using a vector quantizer to generate the classes. The flow regularization method was run twice using different parameters for the smoothness terms, and a strong effect was observed. The higher order bases give an advantage when using rough (large data/smoothness ratio) regularization, as would be expected given that they capture higher spatial frequencies. The low orders gain ground when smoother flows are used, for the simple reason that

more of the work is being performed by the flow generation algorithm, while not affecting the spatial resolution of the flows which they can capture. It has been reported [5] that high resolution flow fields are important for the characterization of facial *action*, a more detailed level of analysis than that of recognizing facial *expression*. Our method performs better in this case, and hence is more extensible towards this more detailed level.

The numbers in Table 4 can be compared with a result of 89.9% obtained from naive (untrained) humans performing the same task [11]. However, the lack of training data in all the experiments discussed here is a limiting factor for our method. Some of the sequences in the *Tulips1* database have only three images, which generate only two flow fields. Given more data, a higher dimensional feature vector could be used. Human subjects have no lack of training data prior to performing recognition tasks.

5 Discussion

We have presented a simple and yet general representation scheme for complex human action based upon a set of orthogonal Zernike polynomials. We have shown that this basis provides good descriptions of flow fields using low dimensional feature vector spaces. It is a holistic approach which is not specific to a given type of motion, and could be simply applied to any type of flow field. A recognition system was built using hidden Markov temporal models, and results were presented which demonstrate the effectiveness of the Zernike basis at characterizing the flow fields generated by human expression with and without significant rigid head motion. Results were compared with those obtained using principal components as a basis, and we found that the universal Zernike basis can outperform the more specific representation, leading to a general method suitable for representing and recognizing arbitrary motions.

Recognizing complex human motions based on flow fields relies on a spatial scale at which to decompose the flow fields, a choice of basis elements to perform the decomposition, and a choice of temporal model. Our research addresses the interplay between these three elements, by investigating the use of Zernike polynomials as basis elements, and HMMs as temporal models. Although we have not directly addressed the issue of scale, the Zernike basis does so implicitly, and is clearly well suited to the task. A future goal of this research is to find optimal trade offs between spatial scale, spatial frequency of the representation, and temporal segmentation, for which the minimum entropy methods of [4] look promising.

References

[1] S. Belkasim, M. Shridhar, and M. Ahmadi. Pattern recognition with moment invariants: A comparative study and new results. *Pattern Recognition*, 24(12):1117–1138, 1991.

[2] M. Black and P. Anandan. The robust estimation of multiple motions: parametric and piecewise smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, January 1996.

[3] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motions. *International Journal of Computer Vision*, 25(1):23–48, 1997.

[4] M. Brand. Pattern discovery via entropy minimization. Technical Report TR-98-21, MERL, Cambridge, MA, October 1998.

[5] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE PAMI*, 21(10):974–989, October 1999.

[6] I. A. Essa and A. O. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proceedings ICCV*, pages 360–367, 1995.

[7] E. Hunter, J.Schlenzig, and R. Jain. Posture estimation in reduced-model gesture input systems. In *International Workshop on Automatic Face- and Gesture-Recognition*, pages 290–295, Zurich, Switzerland, 1995.

[8] A. P. Jr. and W. Rusch. Algorithm for computation of Zernike polynomials expansion coefficients. *Applied Optics*, 28(4):749–754, February 1989.

[9] A. Lanitis, C. Taylor, and T.F.Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE PAMI*, 19(7):743–756, July 1997.

[10] J. J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre*, 1(2), Winter 1998.

[11] J. Luettin and N. A. Thacker. Speechreading using probabilistic models. *CVIU*, 65(2):163–178, February 1997.

[12] C. Morimoto, Y. Yacoob, and L. Davis. Recognition of head gestures using hidden Markov models. In *Proceeding of ICPR*, pages 461–465, Austria, 1996.

[13] J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Toruetzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*. MIT Press, Cambridge, 1995.

[14] M.Turk and A.Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[15] P.N.Belhumeur, J.P.Hespanha, and D.J.Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projections. *IEEE PAMI*, 19(7):711–720, July 1997.

[16] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–296, February 1989.

[17] D. Terzopoulus and K.Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE PAMI*, 15(6):569–579, June 1993.

[18] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. In *CVPR*, pages 70–75, 1994.

[19] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–112, August 1992.