Facial Expression Recognition with Temporal Modeling of Shapes

Suyog Jain, Changbo Hu, J. K. Aggarwal
Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
suyog@cs.utexas.edu, changbo.hu@gmail.com, aggarwaljk@mail.utexas.edu

Abstract

Conditional Random Fields (CRFs) can be used as a discriminative approach for simultaneous sequence segmentation and frame labeling. Latent-Dynamic Conditional Random Fields (LDCRFs) incorporates hidden state variables within CRFs which model sub-structure motion patterns and dynamics between labels. Motivated by the success of LDCRFs in gesture recognition, we propose a framework for automatic facial expression recognition from continuous video sequence by modeling temporal variations within shapes using LDCRFs. We show that the proposed approach outperforms CRFs for recognizing facial expressions. Using Principal Component Analysis (PCA) we study the separability of various expression classes in lower dimension projected spaces. By comparing the performance of CRFs and LDCRFs against that of Support Vector Machines (SVMs), we demonstrate that temporal variations within shapes are crucial in classifying expressions especially for those with a small range of facial motion like anger and sadness. We also show empirically that only using changes in facial appearance over time, without using shape variations, is not sufficient to obtain high performance for facial expression recognition.

1. Introduction

The recognition of facial expressions is a necessary first step for meaningful interactions between humans and computers. A reliable Automatic Facial Expression Recognition (AFER) system will improve the way in which humans interact with machines. Facial Expressions in humans are inherently dynamic in nature, consisting of an onset, peak and an offset phase. The entire event from onset to offset is usually very short in duration, and often the muscle motions on the face are very subtle. This makes the problem of recognizing facial expressions very challenging. In this paper, we consider the problem of recognizing facial expressions from video sequences and formulate it as a sequence labeling problem, where we try to label every frame with the correct facial expression or neutral state.

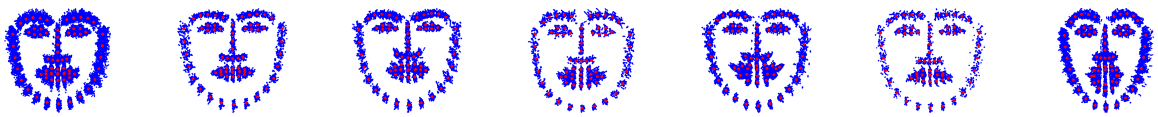
We propose a new approach for recognizing six basic expressions (anger, disgust, fear, happiness, sadness and surprise) along with a neutral state, by modeling temporal dynamics of face shapes. Our approach uses discriminative Latent-Dynamic Conditional random fields (LDCRFs) [11], and we show that incorporating hidden states in traditional Conditional Random Fields (CRFs) [8] model is an effective way to model the subtle changes which happen over time in face shapes. This helps in distinguishing between facial expressions which have large overlapping motion patterns. We also empirically show that classifiers which use temporal variations between shapes outperform those which do not consider this information for the task of facial expression recognition. Finally, we compare the ability to recognize facial expressions using shape variability versus appearance variability and show that variations in shape are much more important than appearance for facial expression recognition from continuous video.

The remainder of the paper is organized as follows: We review the related work in the next section, followed by a description of the features used in our work along with the formulation of both CRFs and LDCRFs. We then present various experiments and compare the results followed by conclusions.

2. Related Work

Facial Expression Recognition has been an area of interest among researchers for several decades. Traditionally, most research for designing such a system has been focused on recognizing expressions from static images. Methods using geometric distances, gabor filter responses and local binary patterns [19, 2, 15] have been successfully applied in this domain. A comprehensive survey of some of these techniques may be found in [12].

Recently, there has been a strong interest in modeling the temporal dynamics of facial expressions to build such systems. The psychological experiments carried out in [1, 3] strongly suggest that modeling facial movements in time is a crucial factor in discriminating between facial expres-



(a) Neutral (b) Anger (c) Disgust (d) Fear (e) Happiness (f) Sadness (g) Surprise

Figure 1. Shapes of various Facial expressions after Generalized Procrustes Analysis. The 5643 face shapes from the dataset grouped by expression labels and aligned using Generalized Procrustes Analysis can be seen here. This figure demonstrates that shapes for some of the expressions like surprise and happiness are easily distinguishable while for other expressions there is a considerable similarity in the face shape. Red dots represent the mean shape.

sions. Here we review some of the work that is closely related with recognizing facial expressions using temporal information.

Cohen et al. [4, 5] use a tree-augmented-naive Bayes classifier (TAN) for continuous videos to learn the correlation between motions of different facial regions and expressions. The authors in [23] propose a method using moment invariants along with Hidden Markov Models (HMMs) to analyze facial expressions. In [22] a volume based appearance descriptor is proposed to recognize facial expressions in image sequences. The authors consider a given image sequence as a whole and classify the entire sequence into one of the expression classes. But for a practical application, a facial expression system must be able to classify images as they come; therefore, a solution which can model the transition between various expressions and label each image continuously is more desirable.

Conditional Random Fields (CRFs), which provide one such solution, were introduced in [8]. These are discriminative models which define a conditional probability p(Y|X) over label sequences Y given a particular observation sequence X. The primary advantage of CRFs over generative models like HMMs [13] comes from the fact that models like HMMs try to define a joint probability distribution p(X,Y) over observation sequences X and their label sequences Y [18]. To make the model computationally feasible, strong independence assumptions among observations are required. In the case of CRFs, the independence assumption has to be made for label sequences and not for observations. Hence, CRFs prove to be more robust in comparison [8].

Sminchisescu et al. [16] have shown the effectiveness of CRFs in recognizing several human motions like walking, running etc. Their method outperforms HMMs and even provides good results for differentiating between subtle motion patterns like normal walk vs. wander walk. The authors in [7] use CRFs to classify facial expressions from image sequences. Their work aims at designing a complete facial expression recognition system but does not provide a detailed analysis on the importance of using temporal information for performing this task. In this paper, we show that the dynamics of shape contain much richer information to recognize expressions in comparison with analyzing each shape individually. We also use CRFs as one of the underlying discriminative classifiers to compare the performance of our proposed approach.

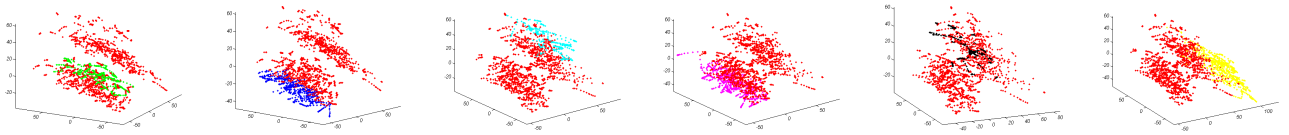
The variants of CRFs which include hidden states have been successfully applied for gesture recognition 1). to label the entire sequence as a whole using Hidden Conditional Random Fields (HCRFs) [21] or 2). to label every frame with the appropriate gesture class using Latent Dynamic Conditional Random Fields (LDCRFs) [11]. It has been shown that these approaches are good at capturing subtle motion patterns using hidden states. Our proposed approach using LDCRFs is more robust in modeling facial expressions as compared to CRFs which shows that capturing subtle facial motion is very essential in differentiating between facial expressions.

3. Methods

In this section we explain the features and classification methods used for our work. Section 3.1 & Section 3.2 gives an overview of the shape and appearance features which were used, while Section 3.3 and Section 3.4 discuss the CRFs and LDCRFs methods which were used to model the temporal variations of these features for facial expression recognition.

3.1. Shape Features

A 2D face shape for our work is represented by a set of 68 landmark points which are basically located around the contours of the eyebrows, eyes, nose, chin, inner lips and outer lips. The distribution of these landmark points on the face can be seen in Figure 1. The task of localizing these landmark points on the face is itself a very challenging problem. Several techniques like Active Appearance Models [6], Piecewise Bezier Volume Deformation (PBVD) [14] etc. have been proposed to solve it. In order to perform a robust shape analysis for different expression shapes, it's important to obtain their true shapes by removing the effects of rigid geometric transformations such as translation, scale and rotation between them.



(a) Anger (Green) (b) Disgust (Blue) (c) Fear (Cyan) (d) Happy (Magenta) (e) Sadness(Black) (f) Surprise(Yellow)

Figure 2. Comparison between first 3 Principal Components (Shape Features) for neutral images with other expressions(Neutral is shown in red in all figures). This figure demonstrates that some neutral images are very close to the expression images in the PCA space. These neutral images mostly correspond to the transition phase from the neutral to the actual expression. Especially for anger and sadness there is a significant overlap which makes these expressions difficult to recognize in presence of neutral images.

We use Generalized Procrustes Analysis for this task [17], which tries to minimize the sum-of-squared distances between landmark points of all the shapes w.r.t. the rigid transformations. In the initial step, the first shape is assumed to be the mean shape. Then, a similarity transform is applied on all the shapes to align them with the mean shape. After this alignment, the mean shape is recomputed by averaging the aligned shapes, and then the process of applying the transforms and mean computation is repeated until the change in the mean shape becomes negligible.

Figure 1 shows the aligned face shapes after Generalized Procrustes Analysis is applied to all the shapes in the dataset. It can be observed that some expressions such as surprise have very distinct shapes while others such as anger and sadness show a certain degree of similarity. After performing the alignment, the true shape which remains gives us a 136 dimensional feature vector.

We apply Principal Component Analysis (PCA) to reduce the dimensionality to 18 by retaining 95% of the variance which forms the input for our classifiers. For practical applications, it's important to consider even the neutral state while designing classifiers for facial expressions. Introducing the neutral state makes the task of recognizing facial expressions more difficult. Many subtle expressions like anger and sadness do not cause much facial movement, therefore they are difficult to differentiate from the neutral state. Also, it is very difficult to make a clear distinction between the end of a neutral state and the onset of an expression even for humans. It makes the task of ground-truth labeling very challenging.

These issues are clearly highlighted in Figure 2 which shows a comparison between first 3 Principal Components of shape features for neutral images with other expressions. For all the expressions, neutral shows some overlap with the actual expressions. These shapes mostly correspond to the transition phase where it's difficult to tell if a shape belongs to the neutral state or to the actual expression. The plots corresponding to anger and sadness show a lot of overlap with neutral shapes in the PCA space, which makes it difficult to recognize these expressions in presence of the neutral shapes.

3.2. Appearance Features

One of the aims of this paper is to experimentally show the importance of temporal variations in shape as compared to the temporal variations in appearance for facial expression recognition.

We use histogram based Uniform Local Binary Pattern (U-LBP) [15] features which are commonly used for facial expression recognition to conduct our experiments. In this method, the LBP operator is applied on a pixel by thresholding its circular neighborhood with the intensity value of the pixel and representing it in binary form (1 if the intensity value of the neighboring pixel is greater than the current pixel, 0 otherwise). The patterns which contain at most two bitwise transitions from 0 to 1 or vice versa are called uniform local binary patterns. It was observed that uniform patterns form the majority of the observed patterns [15], hence to construct the histogram, all unique uniform patterns are binned separately while all non-uniform patterns are assigned to a single bin. We use an 8 pixel neighborhood which gives us a 59 bin histogram.

It has been shown[15] that using a single histogram for the entire image is not a good technique for facial expression recognition, hence the cropped face image is subdivided into 42 regions using a 6 x 7 grid (see Figure 3). Then a separate histogram is computed for each sub-region which gives us a feature vector of length 2478. Principal Component Analysis (PCA) is applied to reduce the dimensionality to 59 by retaining 95% of the variance.

3.3. Conditional Random Fields (CRFs)

CRFs provide a highly discriminative and probabilistic method [8] to model the variation of shapes in time. For comparison with our proposed approach, we use the basic linear chain CRFs for the task of facial expression recognition. (Figure 4).

For notational simplification, we refer the observation sequences (X_1,X_2,X_3,.....,X_T) as X and label sequences (Y_1,Y_2,Y_3,.....,Y_T) as Y for T frames to be labeled. Here

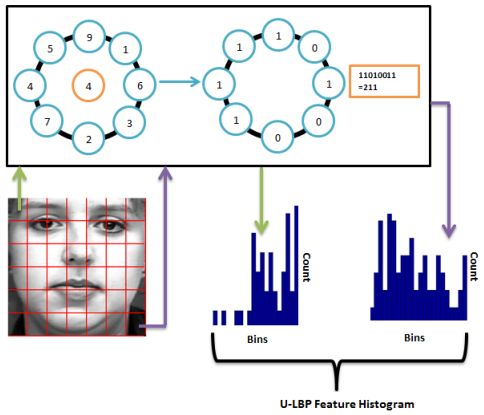


Figure 3. Compuation of Uniform Local Binary Pattern (U-LPB) Histogram. The face image is divided into 6 x 7 grid and then a separate U-LBP histogram is computed within each grid by applying the ULBP operator on every pixel in the grid as shown.

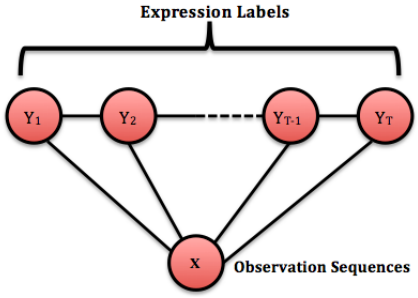


Figure 4. Linear Chain Conditional Random Fields.

each X_i, i $\in (1, 2...T)$ is a random variable representing either the shape or the appearance features and each Y_i, i $\in (1, 2...T)$ is a random variable representing the expression label or neutral state.

A CRF model for T image frames is formulated as follows:

$$P(Y|X;\theta) = \frac{1}{Z(X,\theta)} exp\left(\sum_j \theta_j F_j(Y,X)\right) \quad (1)$$

where,

$$F_j(Y,X) = \sum_{t=1}^{T} f_j(Y_{t-1}, Y_t, X, t) \quad (2)$$

$$Z(X,\theta) = \sum_Y exp\left(\sum_j \theta_j F_j(Y,X)\right) \quad (3)$$

Here, $Z(X,\theta)$ is the normalization factor and each $f_j(Y_{t-1}, Y_t, X, t)$ is either a state function $st_j(Y_t, X, t)$ which evaluates the interaction between features or a transition function $tr_j(Y_{t-1}, Y_t, X, t)$ which models the temporal dependencies among features [20]. Given a set of N labeled training samples, the objective of the training procedure is to estimate the set of weights θ^* which maximizes the conditional log likelihood (i.e. θ^* = argmax_θ L(θ)) by optimizing the conditional log likelihood function given by equation (4).

$$L(\theta) = \sum_{k=1}^{N} \left[\sum_j \theta_j F_j(Y^{(k)}, X^{(k)}) - log\frac{1}{Z(X^{(k)},\theta)}\right] \quad (4)$$

For our work, we use Broyden Fletcher Goldfarb Shanno (BFGS) [10] gradient ascent technique for optimizing the log likelihood function. To classify an unseen test sequence X_{test} we can find the most likely labels Y^* for the sequence using the learned parameters θ^* and equation 1 by Viterbi decoding.

3.4. Latent-Dynamic Conditional Random Fields (LDCRFs)

CRFs provide a strong discriminative framework to model the transitions between facial expressions, but they fail to model the subtle facial motions within an expression which is very important in order to differentiate between similar facial expressions.

In [11] a variant of traditional CRFs known as Latent-Dynamic Conditional Random Fields (LDCRFs) was proposed which captures the subtle motion patterns within a class along with inter-class motion patterns by associating a set of hidden states with each class label. These hidden states can model the internal sub-structure for different facial expressions and contribute in the overall likelihood for recognition. Each hidden state can be treated in a similar manner as CRF, and the overall likelihood can simply be the sum of individual likelihoods from the hidden states.

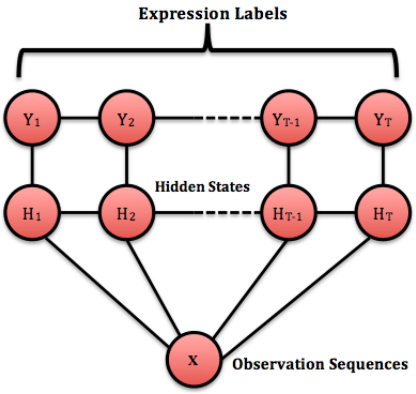


Figure 5. Latent Dynamic Conditional Random Fields.

The LDCRF model uses an additional set of hidden variables H = (H_1,H_2,H_3,.....,H_T) apart from X and Y for every sequence (Figure 5). The model can then be defined over parameters θ as:

$$P(Y|X;\theta) = \sum_{H} P(Y|H,\theta)P(H|X,\theta) \tag{5}$$

The LDCRF model imposes a restriction that sets of hidden states for each class label needs to be disjoint. This implies that for a given class label Y_m the set of possible hidden states H_m is constrained to a subset H_{Ym} of all possible hidden states. This assumption gives the following deterministic relationship between Y and H:

$$P(Y|H,\theta) = \begin{cases} 1 & \forall\, H_m \in H_{Ym} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Hence equation (5) can be refined as:

$$P(Y|X;\theta) = \sum_{H:\forall H_m \in H_{Ym}} P(H|X,\theta) \tag{7}$$

P(H|X,θ) is then defined exactly as P(Y|X,θ) is defined in the previous section.

$$P(H|X;\theta) = \frac{1}{Z(X,\theta)} exp\left(\sum_j \theta_j F_j(H,X)\right) \tag{8}$$

where,

$$F_j(H,X) = \sum_{t=1}^{T} f_j(H_{t-1},H_t,X,t) \tag{9}$$

$$Z(X,\theta) = \sum_{Y} exp\left(\sum_j \theta_j F_j(H,X)\right) \tag{10}$$

As in CRFs, $Z(X,\theta)$ is the normalization factor and each $f_j(H_{t-1},H_t,X,t)$ is either a state function $st_j(H_t,X,t)$ or a transition function $tr_j(H_{t-1},H_t,X,t)$ [20]. The parameter estimation and inference can then be performed in a similar manner as with CRFs. For our work, we use Broyden Fletcher Goldfarb Shanno (BFGS) [10] gradient ascent technique for optimizing the log likelihood function.

4. Experiments & Results

This section gives details about the dataset used for experiments, followed by an overview of the experiments that were conducted. We then present the results of various experiments and compare the facial expression recognition performance for all the techniques from various aspects.

4.1. Overview of the dataset

The experiments for our work were conducted on the Extended Cohn-Kanade Dataset (CK+) [9] which contains 593 sequences from 123 subjects. These are not fixed length sequences and the duration varies from 10 to 60 frames. All the sequences start from the neutral pose to the peak formation of the expression. The locations of facial landmarks are provided along with the dataset. Out of the 593 sequences in the dataset only 309 were labeled as one of the six basic expressions (see [9] for details). Table 1 gives the detailed statistics for the portion of the dataset that was used. The expression onset for all sequences takes place after a certain number of neutral frames; hence we manually label each frame in a sequence to be either neutral or belonging to the expression class. Figure 6 shows an example of one such labeled sequence.

Expression	No. of Sequences	Total No. of Images
Anger	45	1022
Disgust	59	868
Fear	25	546
Happiness	69	1331
Sadness	28	547
Surprise	83	1329
Total	309	5643

Table 1. Overview of the dataset.

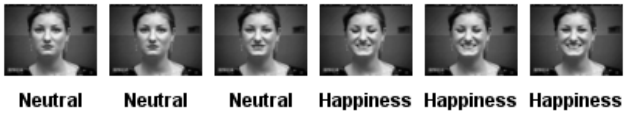


Figure 6. Example of a labeled sequence

4.2. Experiment Details

We perform experiments to show that modeling temporal variation between shapes helps in recognizing those facial expressions which are otherwise difficult to recognize using classifiers which do not model temporal dependencies. For this, we compare the recognition performance of Support Vector Machines (SVMs) classifier trained on the shape features and a baseline method [9] which also uses SVMs against the performance of CRFs and our proposed method of using LDCRFs for recognizing facial expressions. We also show through experiments that modeling variation in shape across time is much more important than modeling variation in appearance across time for recognizing facial expressions. To show this, we train both CRFs and LDCRFs classifiers using appearance features and compare the performance with shape features.

All the experiments were conducted using 4-fold cross-validation and the results were averaged over all the folds. We evaluate the recognition performance in all the experiments for two cases: 6-class (without neutral) and 7-class (with neutral).

4.3. Static vs. Temporal Shape Analysis

For static shape analysis, we use the shape features described in the previous section to train 6-class (without neutral) and 7-class (with neutral) multi-class SVMs. The Radial Basis Function (RBF) kernel along with a grid search to find the best values for C (penalty term) and γ (kernel-width) was used for training the SVMs. For temporal shape analysis, we train a CRF model and validate the regularization term during training. For inference, the model outputs the marginal probabilities for each class label. The class label with the highest probability for each frame is used as the predicted label for that frame. To train the LDCRF model for our proposed approach, the optimal number of hidden states and the regularization term were found using cross-validation during training. It was observed that 5 hidden states give the best results. The training procedure converges in less than 100 iterations.

The results in Table 2 for 6-class classification and the confusion matrices in Table 3, 4 show that Happiness and Surprise are two expressions which are much easier to recognize as compared to other expressions. The recognition performance for both static shape analysis and dynamic shape analysis is very high for these two expressions. It is an intuitive result as these expressions bring a large amount of change in the shape of the face especially the mouth region and thus are relatively easy to recognize. For other expressions such as anger and sadness which do not cause a lot of deformation on the face, the temporal shape modeling performs much better than static shape analysis.

The performance for the SVM based method is very low for sadness - 63.7%. The sadness expression causes very little deformation on the face and hence is very difficult to recognize by looking at shape in isolation. The temporal modeling using CRF improves the performance, but there is a good deal of overlap in the motion pattern for sadness and other expressions, hence the performance is relatively low. The proposed approach using LDCRF successfully models these overlapping patterns using hidden states and captures the subtle differences which improves the accuracy significantly. For anger, disgust & fear both CRF and LDCRF perform in a comparable manner.

Our approach also gives equivalent results for disgust, happiness and surprise to the baseline method and outperforms it significantly for the other expressions.

The case for 7-class classification where we consider neutral frames as well is relatively difficult, because some expressions that have very little facial movement have very similar shapes to the neutral shapes, which makes it hard to discriminate between them. The temporal dynamics between shapes become much more important in this situation and the results in Table 7 show that this is indeed true. As expected, the recognition performance of happiness and surprise is very high for this case, using either the static shape analysis or the dynamic shape analysis. Static shape analysis gives very low performance for fear and sadness expressions. Temporal shape analysis improves the recognition rate significantly for these expressions. The confusion matrix in Table 5 and Table 6 show that neutral frames cause a good deal of confusion with subtle expressions such as anger and sadness and make these expressions difficult to recognize. It can be seen that the proposed LDCRFs based method is capable of better discriminating between the neutral and sadness shapes as compared to both SVMs and CRFs.

	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**
An	**96.4**	1.8	0.0	0.0	0.0	1.8
Di	0.0	**97.6**	0.0	2.4	0.0	0.0
Fe	0.0	0.0	**92.5**	0.0	1.0	6.5
Ha	0.0	0.6	0.0	**99.4**	0.0	0.0
Sa	1.3	0.0	9.2	2.1	**83.8**	3.6
Su	0.0	0.0	1.3	0.0	0.9	**97.9**

Table 3. Confusion Matrix for 6-class classification using CRFs

	An	**Di**	**Fe**	**Ha**	**Sa**	**Su**
An	**97.9**	1.8	0.0	0.3	0.0	0.0
Di	0.0	**97.9**	0.0	2.1	0.0	0.0
Fe	0.0	0.0	**90.5**	0.0	0.4	9.1
Ha	0.0	0.4	0.0	**99.6**	0.0	0.0
Sa	2.8	0.0	2.2	1.5	**90.1**	3.3
Su	0.0	0.0	0.3	0.0	0.9	**98.9**

Table 4. Confusion Matrix for 6-class classification using proposed method based on LDCRFs

	Ne	**An**	**Di**	**Fe**	**Ha**	**Sa**	**Su**
Ne	**72.2**	6.1	2.6	1.6	2.4	10.5	4.5
An	20.6	**73.5**	0.0	0.0	0.0	5.8	0.0
Di	2.7	6.8	**85.6**	0.0	4.8	0.0	0.0
Fe	0.0	0.0	0.0	**94.4**	0.0	5.6	0.0
Ha	0.5	1.0	0.5	0.0	**98.1**	0.0	0.0
Sa	29.1	0.0	0.0	1.3	0.0	**69.6**	0.0
Su	0.9	0.0	0.0	0.0	0.0	0.0	**99.1**

Table 5. Confusion Matrix for 7-class classification using CRFs

It was observed that misclassification usually occurs during the transition phase from one expression to either neutral or to an other expression. We have reported the results

Expression	Baseline [9] (static)	LBP + CRF (dynamic)	LBP + LDCRF (dynamic)	Shape + SVM (static)	Shape + CRF (dynamic)	Shape + LDCRF (dynamic)
Anger	75.00	70.42	76.28	74.70	96.41	97.91
Disgust	94.70	85.54	86.01	87.13	97.60	97.86
Fear	65.20	67.61	80.05	88.77	92.51	90.52
Happiness	100.00	90.90	90.01	98.43	99.41	99.55
Sadness	68.00	60.2	58.68	63.70	83.83	90.08
Surprise	96.00	89.15	87.81	91.67	97.86	98.87
Average	83.15	77.30	79.81	84.06	94.60	95.79

Table 2. Recognition Rates Without Neutral Expression (6-class classification)

	Ne	**An**	**Di**	**Fe**	**Ha**	**Sa**	**Su**
Ne	**73.5**	6.0	1.6	1.9	2.6	9.2	5.2
An	20.6	**76.7**	1.1	0.0	1.6	0.0	0.0
Di	2.7	6.2	**81.5**	0.0	9.6	0.0	0.0
Fe	0.0	0.0	0.0	**94.4**	0.0	4.2	1.4
Ha	0.5	1.0	0.0	0.0	**98.6**	0.0	0.0
Sa	21.5	0.0	0.0	1.3	0.0	**77.2**	0.0
Su	0.9	0.0	0.0	0.0	0.0	0.0	**99.1**

Table 6. Confusion Matrix for 7-class classification using proposed method based on LDCRFs

based on the classification/misclassification of each frame and not considering the dominant expression within a certain window of time. It makes the problem harder because, it's very difficult to label the ground-truth for frames which lie in the transition phase. For practical applications, accurate labeling of every frame may not be required.

4.4. Temporal Shape vs. Appearance in Time

The appearance of the face changes in form of wrinkles and furrows which appear when an expression is exhibited. In this section, we show that in contrast with shape features, temporal variations in appearance alone is not sufficient to recognize facial expressions with high accuracy. Using CRF and LDCRF techniques, we model Uniform Local Binary Pattern (ULBP) based appearance features which are known for their ability to capture these micro patterns (e.g. wrinkles and furrows) on the face and have been successfully used for static facial expression recognition. The results in Table 2 clearly show that except for happiness and sadness the performance is much lower in comparison with the performance of dynamic shape analysis. The interesting thing to note here is that the performance becomes worse on introducing the neutral state (Table 7). The reason for this lies in the fact that for expressions like anger, disgust and sadness the small amount of facial motion does not bring a significant change in the appearance in comparison with the neutral face which results in a considerable overlap of appearance features between them. This makes it very difficult to distinguish these expressions from the neutral state using just the appearance. These experiments show that dynamics of shape and especially ability to capture the subtle motion patterns on the face is very important for robust facial expression recognition.

5. Conclusion

We presented a new approach for facial expression recognition from video sequences using Latent-Dynamic Conditional Random Fields (LDCRFs). The results of our approach show that the expressions such as surprise and happiness which bring significant changes in face shapes are relatively easy to recognize. For other more subtle expressions, classification methods which do not consider the temporal variation between shapes fail to achieve a good recognition rate. Sadness and anger are two of the most difficult expressions to classify especially in the presence of neutral frames. The proposed method was able to perform better as compared to other techniques for these expressions. This shows the importance of modeling small facial motions effectively for recognizing facial expressions.

The experiments show that shape provides much richer information as compared to appearance, and modeling appearance changes in isolation without considering shape changes is not sufficient for robust facial expression recognition. In the future, we want to evaluate the performance of our approach by training and testing it across various datasets and further extending it to work with a live video feed.

6. Acknowledgments

The authors would like to thank the reviewers for their valuable comments which have helped to improve the quality of the paper. This work is partially supported by Instituto de Telecomunicações and the UT Austin/Portugal Program CoLab grant (FCT) UTA-Est/MAI/0009/2009 (2009) supported by the Portuguese government.

Expression	LBP + CRF (dynamic)	LBP + LDCRF (dynamic)	Shape + SVM (static)	Shape + CRF (dynamic)	Shape + LDCRF (dynamic)
Neutral	87.80	85.41	71.32	72.17	73.46
Anger	61.59	62.73	77.09	73.54	76.71
Disgust	65.22	66.80	82.77	85.62	81.51
Fear	47.20	55.43	75.81	94.37	94.37
Happiness	87.84	84.28	96.92	98.06	98.55
Sadness	49.37	51.17	56.15	69.62	77.22
Surprise	91.28	93.76	97.45	99.06	99.06
Average	70.05	71.36	79.64	84.64	85.84

Table 7. Recognition Rates With Neutral Expression (7-class classification)

References

[1] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 2005. 1

[2] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Proc. IEEE Intl Conf. Systems, Man and Cybernetics*, pages 592–597, 2004. 1

[3] J. N. Bassili. Facial motion in the perception of faces and of emotional expression. 1978. 1

[4] I. Cohen, A. Garg, and T. S. Huang. Emotion recognition from facial expressions using multilevel hmm. In *Neural Information Processing Systems*, 2000. 2

[5] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003. 2

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001. 2

[7] A. Kanaujia and D. N. Metaxas. Recognizing facial expressions by tracking feature shapes. In *International Conference on Pattern Recognition*, pages 33–38, 2006. 2

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Fransisco, 2001. Morgan Kaufmann. 1, 2, 3

[9] P. Lucey, J. 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 *CVPR4HB10*, pages 94–101, 2010. 5, 7

[10] A. McCallum. Efficiently inducing features of conditional random fields, 2003. 4, 5

[11] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007. 1, 2, 4

[12] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. 1

[13] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989. 2

[14] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang. Authentic facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 517–522, 2004. 2

[15] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.*, 27:803–816, May 2009. 1, 3

[16] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *In Intl Conf. on Computer Vision*, pages 1808–1815, 2005. 2

[17] M. B. Stegmann and D. D. Gomez. A brief introduction to statistical shape analysis, mar 2002. Images, annotations and data reports are placed in the enclosed zip-file. 3

[18] C. Sutton and A. Mccallum. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006. 2

[19] Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing lower face action units for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, pages 484 – 490, March 2000. 1

[20] H. M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, University of Pennsylvania, Philadelphia, 2004. 4, 5

[21] S. B. Wang, A. Quattoni, L. philippe Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *Computer Vision and Pattern Recognition*, pages 1521–1527, 2006. 2

[22] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:915–928, 2007. 2

[23] Y. Zhu, L. C. D. Silva, and C. C. Ko. Using moment invariants and hmm in facial expression recognition. *Pattern Recognition Letters*, 23(1-3):83–91, 2002. 2