Multiple Cue Integration in Transductive Confidence Machines for Head Pose Classification

Vineeth Balasubramanian, Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing (CUbiC)
School of Computing and Informatics
Arizona State University, USA
vineeth.nb@asu.edu, panch@asu.edu

Shayok Chakraborty
School of Computing and Informatics
Arizona State University, USA
schakr10@asu.edu

Abstract

An important facet of learning in an online setting is the confidence associated with a prediction on a given test data point. In an online learning scenario, it would be expected that the system can increase its confidence of prediction as training data increases. We present a statistical approach in this work to associate a confidence value with a predicted class label in an online learning scenario. Our work is based on the existing work on Transductive Confidence Machines (TCM) [1], which provided a methodology to define a heuristic confidence measure. We applied this approach to the problem of head pose classification from face images, and extended the framework to compute a confidence value when multiple cues are extracted from images to perform classification. Our approach is based on combining the results of multiple hypotheses and obtaining an integrated p-value to validate a single test hypothesis. From our experiments on the widely accepted FERET database, we obtained results which corroborated the significance of confidence measures - particularly, in online learning approaches. We could infer from our results with transductive learning that using confidence measures in online learning could yield significant boosts in the prediction accuracy, which would be very useful in critical pattern recognition applications.

1. Introduction

Transductive inference is a unique form of online learning, where no explicit function is learnt to map input data to output labels. Instead, a classifier in transductive learning provides a prediction on a test data point based on the knowledge it has gained from the data points that it has seen so far - without an explicit function/rule. In traditional learning systems, an inductive approach is followed where given a set of training data $X = \{x_i, i = 1, 2, \ldots, n\}$ and their corresponding labels $Y = \{y_i, i = 1, 2, \ldots, n\}$, the mapping $f : X \rightarrow Y$ between the training data points and the set of labels is learnt as a function. In short, induction can be viewed as reasoning from observed training cases to general rules, which are then applied to the test cases. On the contrary, transductive inference [2] (attributed to Vapnik [3]) is defined as reasoning from observed, specific (training) cases to specific (test) cases. Using transductive inference is motivated by the idea that many a time, a lot of effort (primarily computational) is spent on obtaining an accurate mapping function between the input data and output class labels in a classification problem, which may not actually be necessary. A crude example of transductive learning would be the k-Nearest Neighbor (k-NN) algorithm, where there is no function/rule that is learnt to classify a new data point. Needless to say, the choice of transductive (vs) inductive inference for a particular application will depend on the nature of the application, the nature of data and the nature of the approach taken to solve the problem.

Proedrou et al [1] proposed Transductive Confidence Machines (TCM) as a form of transductive learning for classification, where they simulated the k-NN algorithm using measures of confidence and credibility obtained through their algorithm. While the terms 'confidence' and 'credibility' were defined rather heuristically in their work, the results showed a novel statistical approach to obtaining a measure of confidence associated with a prediction in online classification problems. Traditional approaches provide accuracy of performance on training data, or compute metrics like precision-recall as measures of their performance. However, these measures cannot provide a confidence on the prediction on every test data point uniquely. Probabilistic approaches provide a measure with every prediction - however, most such approaches need the Bayesian assumption; and further, one could argue that

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

there is a subtle difference between probability and confidence. The TCM approach was based on Kolmogorov's theory of randomness [4], and laid only one assumption on the data - that it should be i.i.d (identically independently distributed). The model was initially formulated for classification [1], but also later on extended for regression problems too [5].

The TCM was proposed and applied predominantly on data from the financial domain [6]. In this work, we have adapted the approach to an image-based application - head pose classification using face images. Recent successful approaches to real-world computer vision applications have relied largely on use of multiple classifiers (for example, AdaBoost for face detection [7]). And therefore, while the original method was proposed for a single classifier, we have extended the theoretical framework to support multiple classifiers. We found that this approach gives results comparable to (in fact, better than) the original classifier(s) itself - with the added benefit of a confidence/credibility measure tagged on the final prediction.

In an online learning scenario, it could be expected that as the system learns its model from more training data, even if the predicted value may not be different, the confidence of prediction would get higher with time (and this kind of an observation would not be possible with an offline learning approach). We have performed an analysis on how these confidence measures evolve on a prediction as learning is performed online. Such a system would be useful in an online learning approach, where the learning could be continued until a pre-defined confidence value is obtained. And then, the learning could be suspended until the confidence associated with a prediction falls below a threshold again, thereby waking up the learning process.

To summarize, the main contributions of this work are: (1) the application of online transductive inference to head pose classification using confidence measures, (2) extension of the TCM framework for multiple classifiers to provide a confidence measure on every test data point uniquely, and (3) an analysis of confidence measures for results in online transductive learning. Although proposed with respect to k-NN in this work, this idea can be extended to most other classifiers, as detailed in [4]. We have applied our work to head pose classification and obtained interesting results proving the validity of this approach. To simulate the real-world environment, we obtained our image gallery by running a face detection algorithm on the FERET face image database (more details in Section 4), and used the images obtained from these face detection algorithms as input to our approach. The broader scope of our contributions would include applications of incremental fusion of multimodal data for online classification using transductive inference.

Section 2 provides a brief overview of head pose estimation methods, and discusses the related work on the TCM. In Section 3, we discuss the proposed methodology and how we apply the approach to head pose classification. The experiments, results and the ensuing discussions are presented in Section 4, and we conclude with pointers to promising future work in the final section.

2. Related Work

2.1. Head Pose Estimation: A Brief Survey

The estimation of head pose angle from face images is a challenging problem in computer vision that has received the attention of researchers in the field for more than a decade now. Different approaches have been adopted, and the challenge is intensified with changes in illumination and facial expressions in face images. A survey of different techniques towards head pose estimation has been presented by Balasubramanian et al [8], and the methods adopted so far have been grouped into the categories: *shape-based geometric methods*, *model-based methods*, *appearance-based methods*, *template matching methods* and *dimensionality reduction based approaches*. The interested reader is requested to read [8] for more details on each of these categories of approaches, and their advantages/disadvantages.

In this work, since we intended to focus on the computation of confidence in online transductive learning, we implemented a straight-forward approach to head pose classification, and picked an appearance-based approach. The aspect of related literature that is relevant to our methodology is the nature of image features that have been used in similar applications so far. The image features that have been used can be broadly categorized into two kinds - *appearance-based* features and *facial landmark* features. Examples of appearance-based features include pixel values [9], Laplacian of Gaussian (edge-based) features [8], and Gabor wavelet features [10] [11]; and examples of facial landmark features include approaches where facial features such as the eyes, nose, and lips are localized [12]. In our work, we used appearance-based image features - since facial landmarks are difficult to localize in a consistent manner when there is significant change in head pose angle.

A wide range of classifiers like k-Nearest Neighbors [9], Generalized Regression Neural Networks [8] and Support Vector Machines [10] have been used to approach the learning problem in the current approaches. There have also been efforts where multiple classifiers have been used to learn the

mapping [13] [11]. However, to the best of our knowledge, none of these efforts have used online learning for head pose classification, or obtained a confidence on the final prediction from the classification in this application.

2.2. Online Learning using the Transductive Confidence Machine

A brief overview of the Transductive Confidence Machine (TCM) [1] is presented in this section.

Let the set of training data be defined by the pairs $(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)$, where m is the number of training data points, x_is are the training data points, y_is are the corresponding class labels picked from a finite set of class labels, say: $\{1, 2, \ldots, c\}$. Let D_i^y denote the list of sorted distances between a particular data point x_i and other data points with the same class label, say y. On similar lines, let D_i^{-y} denote the list of sorted distances between x_i and data points with a class label other than y. Also, D_{ij}^y captures the jth shortest distance in the list of sorted distances, D_i^y. Given these notations, every data point in the training set is assigned a value of non-conformity measure (called the 'strangeness' measure in [1]) defined by:

$$\alpha_i = \frac{\sum_{j=1}^{k} D_{ij}^y}{\sum_{j=1}^{k} D_{ij}^{-y}}$$

where k is the number of nearest neighbors. This would be a natural measure for non-conformity, since this measure would increase when the distance of a data point from other points with the same class label increases, or complementarily, when the distances from data points with other class labels become smaller.

Before we continue with the details of the TCM, we review the concept of a p-value in statistics. In statistical hypothesis testing, the p-value is the probability of obtaining a result at least as extreme as a given observation, under the null hypothesis. In simpler terms, if an experiment could yield a result worse than the current result with high probability (assuming the null hypothesis to be true), then the current result corroborates the null hypothesis, and the p-value would be high, and the null hypothesis would be accepted to be true. For more insights and examples, please refer [14].

When a new test data point, say x_{test}, enters the system, a null hypothesis is assumed - for instance, let x_{test} belong to the class label y_{test}, where $y_{test} \in \{1, 2, \ldots, c\}$. The non-conformity measures of all the data points in the system so far are re-computed with the assumption of the null hypothesis. To validate the proposed null hypothesis, a p-value is defined as follows:

$$p_{y_{test}}(\alpha_{test}) = \frac{count\{i : \alpha_i \geq \alpha_{test}\}}{m+1}$$

where α_{test} is the non-conformity measure of x_{test}, assuming it is assigned the class label y_{test}. It is to be noted that this definition supports the inference that the p-value would be highest when all non-conformity measures in the training data are higher than that of the test data point i.e. the test data point conforms to the proposed class label better than most training data points (or even, all of them) - and hence, the null hypothesis could be accepted. This process is repeated with the null hypothesis supporting each of the class labels, and the highest of the p-values (termed 'credibility') is used to decide the actual class label assigned to x_{test}. Since the best chance that x_{test} belongs to another class label would be given by the second highest p-value, the 'confidence' of prediction is heuristically defined as: $1-$ the second largest p-value. For more details on how the defined expression qualifies as a p-value and its basis on the Kolmogorov complexity and Martin-Lof test of randomness, please refer [4].

While k-NN is traditionally known to be a 'lazy' learner, the TCM is more involved in terms of computation, and is grouped in the family of online learning techniques. Sample applications of the TCM include financial data analysis [6] and medical diagnostics [15].

3. Proposed Methodology: Integrating Multiple Cues for Head Pose Classification

In this work, we propose to perform multiple cue head pose classification from face images using online transductive inference, and attach a measure of confidence to the final predicted class label. While existing techniques try to obtain an accurate pose angle estimate from the image or classify the face images into discretized bins of pose angle between -90° and $+90^\circ$, our work is targeted towards classifying face images coarsely into the classes: 'Looking straight', 'Looking to the left', 'Looking to the right', 'Looking to the far left', 'Looking to the far right'. The possible ambiguity in the distinction between these class labels is taken into account in our discussion of results (Section 4.2). We intend to apply this work towards a wearable camera-based system for individuals who are blind/visually impaired to enhance their social interaction in daily life (with a setup similar to Krishna et al [16]). We believe that this formulation of the problem would serve our purpose of helping such a user to stay informed of the relative direction in which an individual in the scene is looking (Figure 1 shows an application scenario).

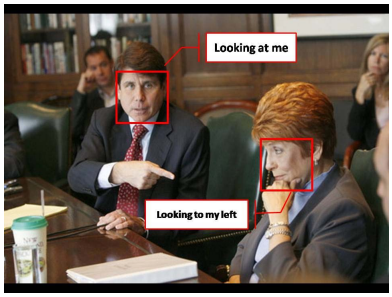


Figure 1. A sample application scenario for the proposed head pose classification system

The input to our system is the face portion extracted from regular images. This is done using one of the real-time face detection algorithms [7] [17], which have been known to be efficient in recent years. In our initial experiments, as would be expected, we found that it was difficult to obtain good results consistently for head pose classification from face images with only one set of features extracted from these images. Hence, we adopted the multiple cue approach to the problem. In this work, we considered three kinds of appearance-based features extracted from the face images: gray-scale pixel values, edge-based features (going by the intuition that edges would be more decisive than textures in pose estimation [8]), and Gabor wavelet features captured at different scales and orientations [11].

We use the Transductive Confidence Machine approach to provide a prediction similar to the k-NN classifier, along with the confidence measure. The prediction happens online, and the non-conformity measures of all data points are updated when a test data point enters the system. The existing TCM framework assumes that only one kind of experiment (based on a single image feature) is studied to accept/reject the null hypothesis. We propose a novel framework based on the TCM for multiple independent tests (each based on one image feature), which validate a single combined hypothesis. We believe that this will permit extensibility of the framework and also, help study the effect of each test in such a framework independently from the other.

In statistical terms, when we consider multiple observations for analysis under a single null hypothesis, we intend to assess the overall significance of a body of experiments possibly containing a mixture of significant and nonsignificant results, with each experiment having its own p-value [18]. In our case, our null hypothesis would be stated as: the test image x_{test} that has entered the system is assigned the class label, y_{test}. We obtain a unique p-value for this case, and similarly obtain p-values for each of the other image features independently - which is a measure of the statistical significance of each of these observations in validating the null hypothesis. We now need to obtain a single integrated p-value from combining these p-values to accept/reject our null hypothesis.

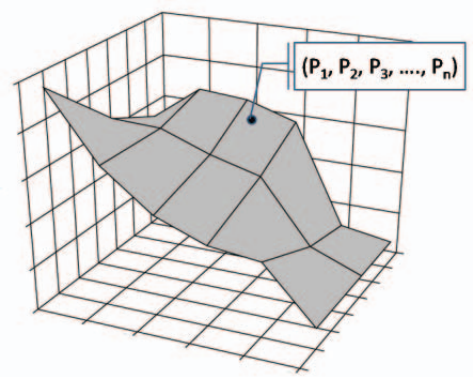


Figure 2. A surface of points with the same probability as the point $(P_1, P_2, P_3, \ldots, P_n)$ representing the p-values P_i of each of the n experiments

A p-value is assumed to be a uniformly distributed random variable on the interval $[0, 1]$. For n experiments or analyses, one can create an n-dimensional unit hypercube and plot the point $(P_1, P_2, P_3, \ldots, P_n)$ representing the p-values P_i of each of the n experiments. A surface of points with the same probability as this point can then be established (Figure 2). Since the p-values are independent probabilities (under the null hypothesis), the individual probabilities can be multiplied to give the probability of obtaining this set of p-values. The set of points whose probability is equal to that of the given set of p-values is then the hypersurface:

$$(x_1 \times x_2 \times x_3 \times \cdots \times x_n) = k$$

where $k = (P_1 \times P_2 \times P_3 \cdots \times P_n)$, the product of the given set of p-values.

By definition of a p-value, we need the probability of getting a vector of p-values as extreme or more extreme (under the null hypothesis) than the given set of p-values. Therefore, we need to find the volume under this hypersurface. Because p-values are uniformly distributed random variables, and because the total volume of the cube equals 1, the volume under the surface directly gives the probability of obtaining a set of p-values as extreme or more extreme than the given set. The volume integral depends only on k (the product of the given set of p-values) and n, the number of p-values under consideration. The overall significance level, for the case of two p-values, is then given by [18]:

$$k - k \ln k$$

And for n tests, the combined significance level is given by:

$$k \sum_{i=0}^{n-1} \frac{(-\ln k)^i}{i!}$$

We use this formula to combine the results from each of our features to obtain a final p-value which decides the predicted class label, and computes a confidence measure value.

4. Experimentation, Results and Discussion

4.1. Experiment Setup

In order to address the application motivating our efforts (as discussed in Section 3), an experiment that would ideally reflect the real-world setting would be a wearable camera that captures images of scenes with faces, coupled with a face detection algorithm that extracts the face portions from these scenes for further analysis. To obtain face images that are reasonably close to such a real-world setting for our analysis, a real-time face detection algorithm based on patch classifiers [17] was applied on images from the FERET face database [19]. In our initial experiments, we found that the work of Kienzle et al [17] performed better in detecting faces with a large range of pose angles, as compared to the popular Viola-Jones face detection algorithm [7]. Samples of face portions extracted from the FERET database which were used for our analysis, are shown in Figure 3.

(a) (b)

(c) (d)

Figure 3. Face images from the FERET database (on the left) and the corresponding extracted face portions (right) used in our analysis

From these face portions, three different features were extracted:

1. The pixel intensity values.
2. The edges were extracted out of the image by thresholding the magnitude of the gradient computed from the vertical and horizontal Sobel filters, δ_y and δ_x respectively. The orientation of each of these edge pixels is computed as:

   $$\theta = \tanh(\frac{\delta_y}{\delta_x})$$

   Subsequently, a histogram of the orientations of these pixels is constructed with bins spanning the interval $[-180°, +180°]$. Initial experiments were carried out with 6, 8 and 12 bins with a regular k-NN classifier to study the performance, and a histogram with 12 bins was found to be most suitable.
3. Gabor wavelet features at three different scales ($\{1, 2, 4\}$) and three different orientations ($\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$) were extracted from the image, and concatenated.

The ground truth for the class labels of these images was collected by manually labeling each of the images into the classes: 'Looking straight', 'Looking to the left', 'Looking to the right', 'Looking to the far left', 'Looking to the far right'. This was carried out using the FERET file nomenclature, where the last two letters of the file name of each FERET image indicates the pose angle of the face in the image [20]. With this knowledge, we used the following guideline to label the images (a similar rule with negative angle values applies for the left):

- *Frontal:* $0\,°$
- *Right:* $> 0\,°$ and $\leq 45\,°$
- *Far right:* $> 45\,°$ and $\leq 90\,°$

While there is the possibility for subjectivity in the kind of class labels we have chosen, we factor this ambiguity in our analysis of results.

Different experiments were conducted to study the performance of the proposed methodology. Note that in each of these experiments, the phrase 'images from a database' is to be understood as the face portions of the images extracted using the face detection algorithm, as explained earlier.

- **Experiment 1:** As a baseline algorithm, k-NN was applied on each of these three features with 3415 images from the FERET database. About 1709 images were used as training images, and the remaining 1706 images were used for testing. These sets of images were picked from the database using a randomized selection process, and this random seed was retained for other experiments to allow comparison.

- **Experiment 2:** The proposed method of integrating p-values for multiple cues in the TCM framework was implemented, and tested with the same set of images as in Experiment 1.
- **Experiment 3:** To study the effect of online learning, Experiment 2 was again carried out - but this time, starting with a set of only 100 training images from the FERET database. The set of training images was gradually increased in steps of 50, and the predicted result, credibility and confidence measures were noted - to be studied for how these evolve with online learning. We terminated this process when all the 1709 training images had been used. It would be expected that as the number of training images increase, the system would be able to predict a result with higher confidence.

To measure the performance of these experiments, the criteria that were used for evaluation are listed below:

- Firstly, the predicted class label was compared with the ground truth to infer a preliminary level of accuracy. In case of face images that were misclassified with an adjacent class label (for example, 'looking to the left' was misclassified as 'looking to the far left'), we noted these cases separately since these would not be strict misclassifications.
- In the experiments with the proposed approach, the accuracy obtained with different confidence value thresholds were noted, and the trend of the accuracy obtained at various confidence levels was studied.
- The effect of online learning was evaluated by studying the evolution of the confidence measures as more training images were added to the system.

The results of these experiments are discussed in the next section.

4.2. Results and Discussion

The results of Experiment 1 for k-NN (k = 10) with the three features are summarized in Table 1. As mentioned earlier, 1706 face images from the FERET database were tested. The first column in Table 1 shows the accuracy i.e. the number of correct classifications with the test images. Ideally, the sum of the portions in the 2nd and 3rd columns would report the misclassifications of the experiment. However, to account for the ambiguity in the class labels, the misclassifications due to adjacent class labels are noted separately. An example of an ambiguous misclassification is presented in Figure 4. As mentioned before, these cannot be termed strict misclassifications, for the predicted class label has a reasonable amount of correctness. As Table 1 shows, the actual number of misclassifications is very small.

	Classifications		
Feature used	Correct	Incorrect	Incorrect but adjacent
Gray-scale pixel values	71.39%	1.83%	26.78%
Edge orientation histogram	52.23%	14.30%	33.47%
Gabor wavelet features	68.87%	1.82%	29.31%

Table 1. Results of k-NN with three different features on test images from the FERET database

Figure 4. **Incorrect but adjacent:** An image from our test set, which was manually labeled as 'Far left', but was classified as 'Left'

The preliminary accuracy of the results of Experiment 2 are noted in Table 2. As is evident from Tables 2 and 1, the accuracy levels obtained from our approach matches (marginally better than) the k-NN approach, as we intended. However, the more interesting part of the results of this experiment is captured in Figure 5 and Table 3. In these results, only predictions that satisfy a specified confidence threshold criterion are considered for evaluation. For results that fall below this specified threshold, if the application demands a high accuracy of prediction, the system can be made to be 'non-committal' about its prediction. From Figure 5, the accuracy grows exponentially as we consider the results that have a confidence score above the specified threshold values. It is also interesting to note that the count of ambiguous predictions is reduced (Note columns 2 and 3 on Table 3). While this result seems intuitive at first glance, it is worthwhile noting that this result establishes that the formulation of confidence in our framework satisfies the intuitive notion of confidence that we hold.

Classifications		
Correct	Incorrect	Incorrect but adjacent
71.45%	1.82%	26.73%

Table 2. Results of our approach using integration of multiple cues in TCM

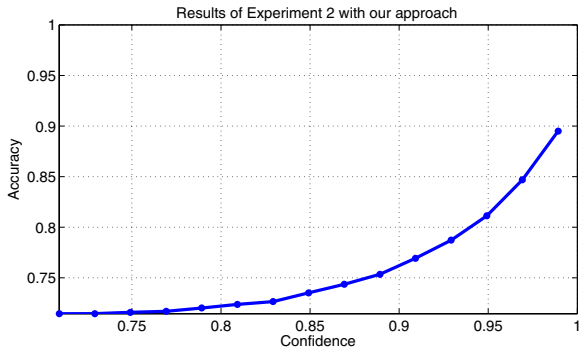
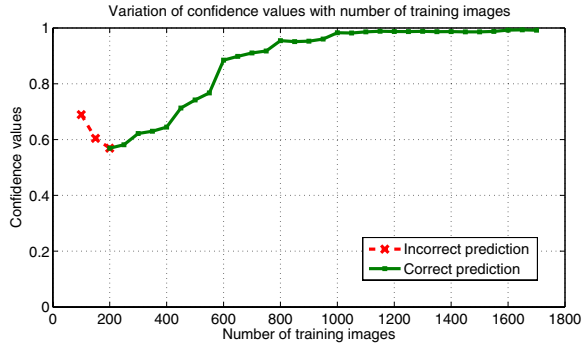


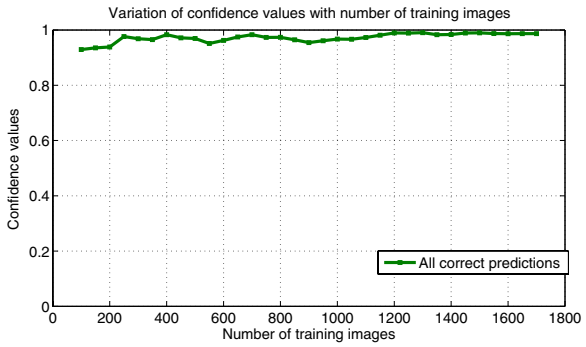
Figure 5. Plot of accuracy of our approach (vs) the confidence measure threshold value. Note that as the confidence threshold is set higher, the accuracy almost increases exponentially

	Classifications		
Confidence threshold	Correct	Incorrect	Incorrect but adjacent
0.70	71.45%	1.82%	26.73%
0.75	71.59%	1.83%	26.58%
0.80	72.00%	1.83%	26.17%
0.85	73.57%	1.65%	24.79%
0.90	75.81%	1.08%	23.12%
0.95	81.18%	0.56%	18.26%
0.98	86.11%	0.84%	13.05%

Table 3. Results of Experiment 2 using our approach, where the accuracy has been measured with different confidence threshold values



(a)



(b)

Figure 6. The evolution of confidence with online training as the number of training images are increased. These two sub-images are representative of the trends that were observed with most test images.

The evolution of the confidence measure with online training (Experiment 3) is traced in Figure 6. Many test data points (about 60 of them with varied parameters for k in k-NN) were classified progressively with increasing amounts of training data (ranging from 100 to 1706). With the results obtained, there were two major trends that were observed of the evolution of confidence with online learning. Figure 6(a) presents a trend, where our classifier initially predicted the wrong label; but as more training was performed, the confidence started dipping, before the correct label was predicted and the confidence measure continued to increase to a very high value when the training was terminated. In case of Figure 6(b), the trend observed was more trivial, and the classifier predicted the correct label with a high confidence right at the beginning, and this confidence continued to increase very slowly over the duration of the online learning process.

5. Conclusions and Future Work

In this paper, we have presented an online transductive learning approach to head pose classification with multiple cues, which can provide an estimate of confidence of the final predicted class label. Using Transductive Confidence Machines as our basis, we have extended the framework to include multiple classifiers (multiple tests, in statistical terms) to arrive at a confidence (p-value) for the final prediction. Our results show that the approach has promise, and the results obtained with our confidence measures exhibit that such a framework would be extremely useful in critical applications, where it would be necessary to make a prediction with high confidence and not make a prediction at all when the system is not very confident of a prediction.

The use of multiple cues in our work allows the approach to be extended as a framework for other learning algorithms with multiple classifiers (such as AdaBoost and classifier ensembles), and applications involving multiple cues (such as multimodal fusion). Also, this work opens up the possibility of using the confidence value as a measure to decide how long the online training process needs to be continued.

One of the seeming limitations of this work is the number of data points that can be classified with a high confidence measure value. As Figure 7 illustrates, and quite obviously, the count of data points that can be classified with

an increasing confidence threshold decreases, as the threshold increases. For example, the number of data points on which the system would be 'non-committal' is high when the confidence threshold is very high ($> 95\%$). However, this could be useful in critical applications, where it may be better to be absolutely sure about few, than be approximately correct for many. Also, more than 70% of the total data points are still classified with at least 90% confidence, which is reasonably good.

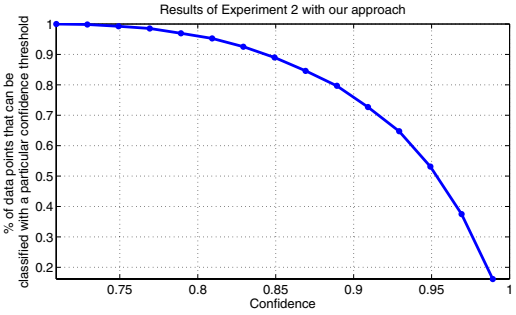


Figure 7. Plot of the portion of test data points that are classified with a particular confidence threshold. As would be expected, this count decreases with increasing confidence thresholds.

The current work has been carried out with k-nearest neighbors to verify the validity of the proposed approach. In our future work, we intend to include Support Vector Machine based classifiers in this framework to see how it scales to other classifiers. We also intend to work on a methodology to evaluate a value of confidence that is computed using user feedback and human-in-the-loop approaches. In another dimension, we plan to adopt a fuzzy formulation of the problem to resolve the ambiguity in the class labels in the proposed system.

References

[1] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman, "Transductive confidence machines for pattern recognition," in *Proceedings of the 13th European Conference on Machine Learning*. Springer-Verlag, 2002, pp. 381–390.

[2] "Transduction (machine learning)," http://en.wikipedia.org/wiki/Transduction_machine_learning, last Accessed: Mar 25, 2008.

[3] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, Sep. 1998.

[4] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 1st ed. Springer, Mar. 2005.

[5] I. Nouretdinov, T. Melluish, and V. Vovk, "Ridge regressioon confidence machine," in *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 385–392.

[6] "Confidence machine," http://www.eghamtech.com/confident_trader.html, last Accessed: Mar 25, 2008.

[7] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, pp. 137–154, 2004.

[8] V. N. Balasubramanian, S. Krishna, and S. Panchanathan, "Person-independent head pose estimation using biased manifold embedding," *EURASIP Journal on Advances in Signal Processing*, 2008.

[9] Y. Fu and T. S. Huang, "Graph embedded analysis for head pose estimation," in *7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, 2006.

[10] X. Ge, J. Yang, T. Zhang, H. Wang, and C. Du, "Three-dimensional face pose estimation based on novel non-linear discriminant representation," *Optical Engineering Letters (SPIE)*, vol. 45, no. 9, 2006.

[11] S. Ba and J. Odobez, "A probabilistic framework for joint head tracking and pose estimation," in *IEEE International Conference on Pattern Recognition (ICPR04)*, 2004, pp. 264–267.

[12] K. Kinoshita, M. A. Yong, L. A. O. Shihong, and M. Kawade, "A fast and robust facial points localization and pose estimation system," 2005.

[13] S. Gundimada and V. Asari, "An improved snow based classification technique for head pose estimation and face detection," in *Proceedings of 34th Applied Imagery and Pattern Recognition Workshop (AIPR05)*, Washington, DC, 2005.

[14] "What is p-value," http://en.wikipedia.org/wiki/P-value, last Accessed: Mar 25, 2008.

[15] Luo, Bellotti, and Gammerman, *Qualified Predictions for Proteomics Pattern Diagnostics with Confidence Machines*, 2004, pp. 46–51. [Online]. Available: http://www.springerlink.com/content/h0ucbd40r7tw40n7

[16] S. Krishna, G. Little, J. Black, and S. Panchanathan, "A wearable face recognition system for individuals with visual impairments," in *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*. Baltimore, MD, USA: ACM, 2005, pp. 106–113.

[17] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf, "Face detection - efficient and rank deficient," *Advances in Neural Information Processing Systems*, 2005.

[18] L. Jost, "Combining significance levels from multiple experiments or analyses," last Accessed: Mar 22, 2008.

[19] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1104, 2000.

[20] "Feret nomenclature," http://www.itl.nist.gov/iad/humanid/feret/feret_master.html, last Accessed: Mar 25, 2008.