Social Signal Processing using sequence of hand gestures to predict explicable social behavior

Soumajyoti Sarkar, Avik Pal, Jaya Sil

Indian Institute of Engineering Science and Technology, Shibpur, India

I. Introduction

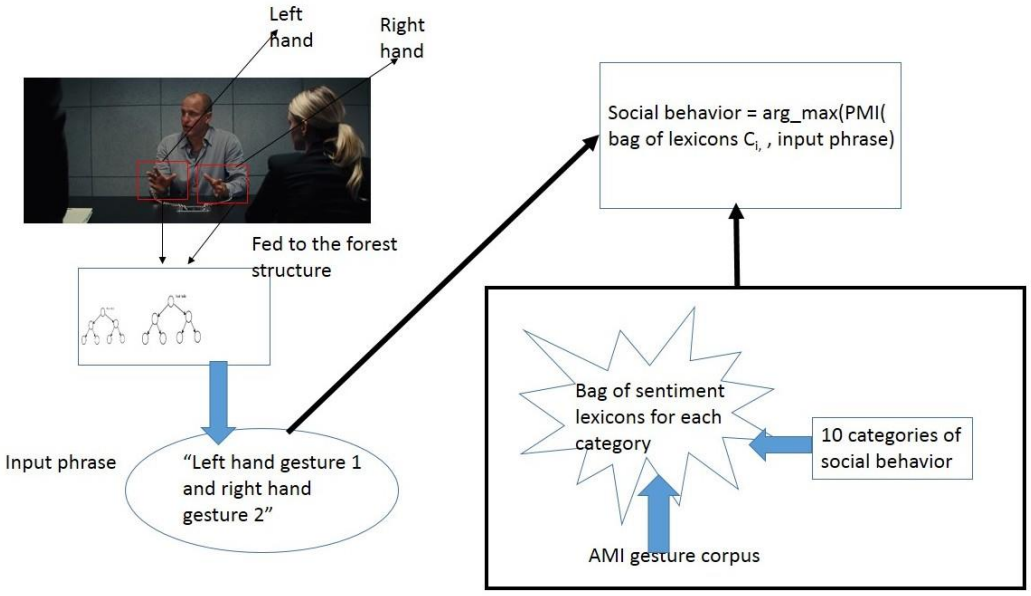
Social Signal Processing is a new domain which attempts to enable machines to replicate human intelligence by analyzing and understanding social interactions between people. The ability to understand the way people interact and recognizing the social behaviors emerge from the social signals is the very essence of social intelligence. Therefore, the social signal processing research integrates social psychology and pattern recognition in order to assess social signals and develop techniques that would empower machines to gauge human social behavior. The term Social Signal Processing was first termed by Pentland. Social signal processing differs from the generic affective computing by means of analyzing specific social signals like hand gestures, facial expressions, head orientation, body posture etc. The signals are modeled to understand the human behavior considering surroundings of the person. Furthermore, social signals differ from the regular social behavior, which are shorter in duration compared to social behavior like politeness, empathy and take hours to capture.

The importance of studying social signal processing is evident in a lot of situations where studying the user behavior is important. For example, in a classroom environment, it is essential for a human teacher to capture the students' reactions and respond accordingly as needed. Therefore, such non-verbal behavioral cues are a regular source of social signals, and are very important in human computer interaction scenarios. The term behavioral cue indicates the psychological activity arising out of the social signals form the basis of social signal processing. As indicated in [], behavioral cues which are constituted by affective states like fear, joy, emblems like thumbs up, manipulators like lip biting, illustrators like finger point raised as well as regulators like smiles, may include both verbal and non-verbal communication. In that same study, it was concluded that non-verbal cues actually dominate the role in assessing the perception of social situations.

Hand gestures are important in modeling social interactions and in most cases used to regulate interactions, to communicate a specific meaning, to punctuate a discourse or to greet. In some cases, the gestures are performed unconsciously and they are interesting from an SSP point of view because they convey honest information. Hands are the most regularly used objects for conveying emotions and ideas while communicating verbally and assist people in expressing themselves properly. The role of Social Signal Processing using hand gestures can substitute semantic analysis of hand gestures with that of verbal cues and can actually escalate the intent of the speaker. Hand gestures are very important in environments of meetings, interviews, debating since they are reflexes used to spread a certain meaning to the listeners in that environment. Hand gestures as described in [] belong to mainly three categories: (i) communicative, which are more of a short stimuli after a prolonged conversation like nervous impacts,

(ii) Co-verbal gestures are communicated in sync with other gestures like speech, face movements and (iii) Improvised gestures, which are completely made up to communicate something.

In the paper we have used coverbal gestures to validate the results on the social behavior. To the best of our knowledge, this is the first attempt to detect social behaviors using hand gestures as social signals. Our proposed method focuses on creating a scalable model to transform communicative hand gestures depicted in a sequence of images into an interpreter consisting of patterns of social behavior that the person is likely to display. For understanding social interactions from hand gestures has been divided into four phases: (i) hand region segmentation using YCbCr colour model, and its view invariant representation using the proposed curvature scale space distribution (CSSD) model, (ii) feature extraction using pyramid histogram of oriented gradient (PHOG) descriptors, (iii) learning the semantics of hand gestures using forest data structure and (iv) augmenting the semantic text of hand gestures into behavioral actions by preparing bag of semantic lexicons of a training corpus.



II. Hand Region Segmentation

Most of the complete hand interactive systems can be considered to be comprised of three layers: detection, tracking and recognition. The detection layer is responsible for defining and extracting visual features that can be attributed to the presence of hands in the field of view of the camera(s). The tracking layer is responsible for performing temporal data association between successive image frames. In the recognition layer the meaning of hand movements are interpreted.

In the proposed detection layer we segment out the hand regions of the person whose social behavior has to be assessed using the concept of curvature scale space matching. The algorithm has been performed on an input image window in three stages: first the skin regions are segmented using a global threshold technique. Then a curvature scale space (CSS) image is created considering the largest contour of the segmented skin region. Finally, a novel approach using statistical measure has been applied to match the input CSS image and the set of previously stored model CSS images. The idea behind our work has been an extension of [] where curvature scale space matching is used to find similarity between shapes. The robustness of that work lies in the fact that CSS matching takes into account scale, rotation as well as affine invariant features in order to apply to the spatio-temporal time series. This concept applies well to the hand regions since a particular hand gesture can be in different orientations and so a view-invariant representation of the hand is required for matching, an important property handled efficiently by CSS images. In the paper we propose a new statistical approach to match the CSS image of input image window with the stored CSS model images by considering their probability distributions and measuring the differences of the distributions by taking into account both the local and global characteristics of the distributions. This is a major contribution is reported in the paper to overcome the limitation of the matching algorithm [6].

The following steps are executed to match the hand gesture in question with the stored images.

(i) *Segmenting the skin region:* A color balancing gray world algorithm [8] has been used to remove the color casts from the background of the hand region so as to fine-tune the pixels for better threshold. We extract the skin pixels from the image window by selecting a threshold value using YCbCr color model. The color image is finally converted into a binary threshold image. From the threshold image, the one with maximum contour length boundary is selected for generating CSS image. We assume that in a window, the hand region can be identified as the object with the largest boundary out of all skin pixel regions in that window, as shown in Fig.1. The corresponding CSS image and CSSD image are shown in Fig.2 and Fig.3 respectively. Finally, Fig.4 shows result of the CSSD image based matching algorithm.

Fig.1: Extracting the biggest contour from the frame

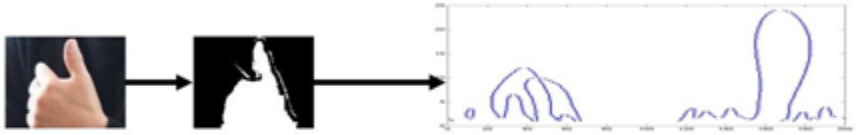


Fig.2: CSS Image Computation

Fig.3: CSS Image to Probability Distribution

CSSD matching:indicate which one is matched to whom mentioning in the figure.

III Feature Extraction

In the paper we have considered assessing the social behavior of a person with a single instance within the frame, not multiple people in a single frame. Once we have extracted the hand regions we perform the following steps.

(i) If there are more than two hand regions are detected, the first frame is labeled as a left hand and the other as a right hand. WHY?

(ii) Else if only one hand region is detected, we label the first frame as left and the second as right.

In the second phase of the work a model has been proposed to extract features from the segmented hand region using the Histogram of Oriented Gradients (HOG) as feature descriptors to represent the hand gestures. HOG descriptors use locally normalized histogram of gradient orientations within a dense overlapping grid, most relevant since the boundaries of hand regions of hand gestures exhibiting lots of alignments. The gradient histogram used in HOG is a form of "quantization", where in this case we are reducing 64 vectors with 2 components each down to a string of just 9 values (the magnitudes of each

bin). Compressing the feature descriptor may be important for the performance of the classifier, but the main intent here is actually to generalize the contents of the 8×8 cell. The histogram bins allow for some play in the angles of the gradients, and certainly in their positions (the histogram doesn't encode where each gradient is within the cell, it only encodes the "distribution" of gradients within the cell). In our case we used 9 bins for computing the histogram. In Fig.5 we show the output of the HOG descriptors on hand gestures.

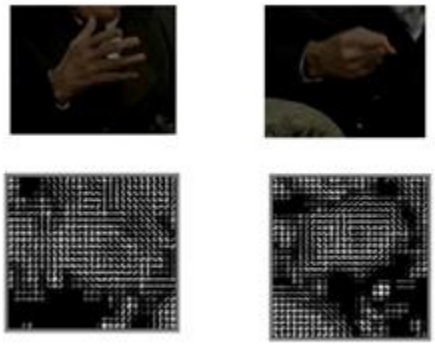


Fig.5: Hand Gestures and HOG Representation

However, using of HOG descriptor is not enough because it does not keep information about the spatial layout of the hand gesture within the frame. In order to determine the relative position and orientation of the hand in one frame with respect to its predecessor within a sequence of frames, it is essential to capture the spatial layout of the hand gesture. Secondly, hand regions vary in size due to the orientation, which is not dealt by the HOG descriptors, shown in Fig.6.

Fig.6: The size of the Hand Region in a Frame can Vary

Therefore, to keep track of similar gestures, we need to include information at different levels at which the HOG descriptors are computed, as shown in Fig.7. Pyramid HOG (PHOG) descriptor [] representing shape of an object with a spatial pyramid kernel has been invoked in the paper to overcome the limitations of HOG descriptors. In PHOG the shape of an object is represented in the form of edges, replacing the visual words concept used in [14]. For each level of the pyramid, a weight is associated with the spatial histograms, which accounts the geometric variability in some hand gestures and invariability in case of others. We use the concept of PHOG in a slightly different way where information about the geometric layout of the hand region in a sequence of gestures is recorded as the spatial difference of the hand regions between two successive frames.

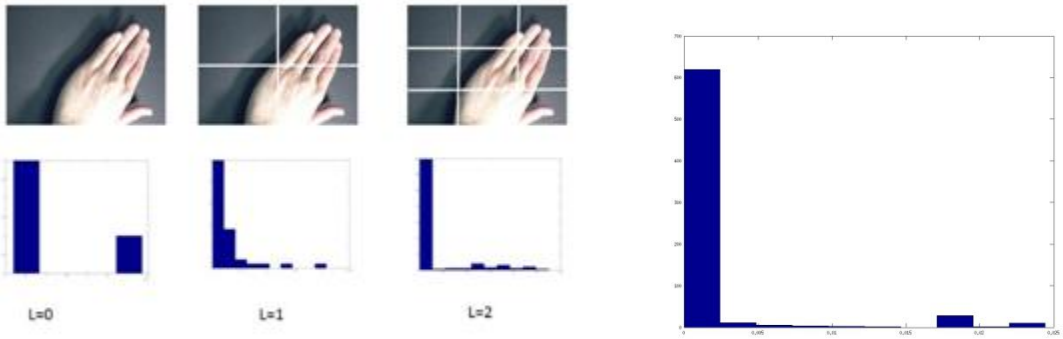


Fig.7: HOG Descriptors at three different levels

We use PHOG in order to capture the spatial difference between two consecutive frames of a single hand gesture and marked as an identifier for measuring differences between two differing gestures.

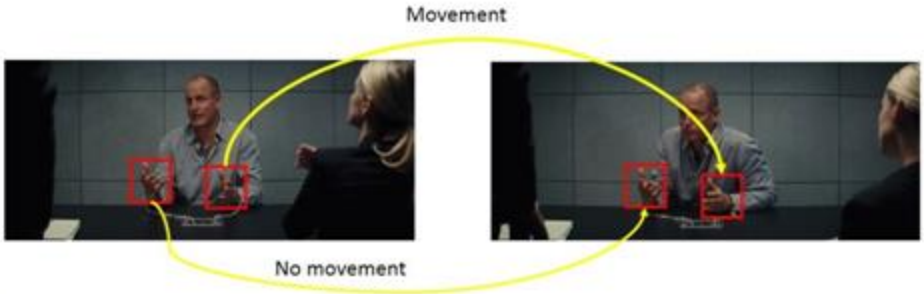


Fig.8: The above images capture two successive frames in a movie sequence. Notice that there isn't any spatial difference in the left hand between the two frames but there has been movement in the right hand. PHOG helps in capturing this spatial difference with respect to the right hand in this case.

The PHOG descriptors score better over other similarity distance measures because it can able to capture slight variation of hand regions between two consecutive frames of a hand gesture in terms of spatial orientation. There might be a huge difference between the last and the first frame of a hand gesture, but between two consecutive frames of a sequence there exist only slight variation (see Fig.8) in terms of the descriptors and used for representing the key points in the hand gesture.

IV. Learning of the Patterns

The hand gestures are learnt by investigating several frames of hand sequences represented by tree structure. Tree representation is used to model the hand gestures which often start with a similar hand orientation and tend to change over the next few frames in the sequence. The concept is primarily based on the Semantic texton forests (STF) [25], ensembles of randomized decision trees which model input video patches or feature points into semantic textons. STFs provide a powerful discriminative codebook using multiple decision trees and the training process of STFs is similar to that of random forests. At each split (?) node, candidate split functions are generated randomly, and the one that maximizes the information gain ratio (between which edges) is chosen. Our model of tree based structure is based on the same fact but at a higher level concept where we consider each node in the semantic texton forest represents a frame consisting of feature points defined using PHOG descriptor. The tree is thus constructed describes a sequence following the path from the root node to the leaf.

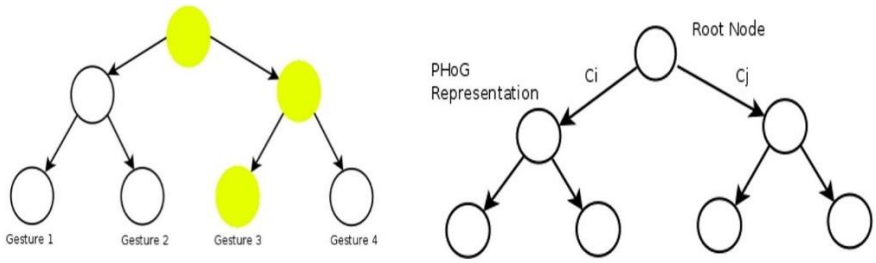
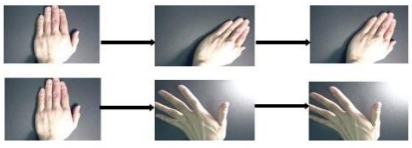


Fig.

Describe the tree, especially the nodes with different colour.



Two sequences of gestures starting with the same frame.

The above images show two sequences of hand gestures captured from two different movies. While both of the gestures have the same starting hand gesture, the first sequence depicts a light mood conversation expressing some statement, the second sequence depicts a heated conversation expressing a resolute statement.

We propose a computationally efficient matching and retrieval method for classifying new sequences of hand as a particular gesture, based on the actions previously learnt and stored using the training data set.

The purpose of redundancy is not clear.

In order to reduce the size of the training data set for on-line application, we incorporate redundancies within the training set by modeling the gestures as an incremental tree as shown below.

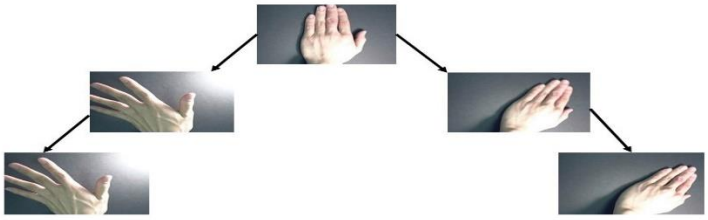


Fig.

Fig. demonstrates that starting with a common frame we can arrive at multiple gestures represented by a common parent. The gestures can then be modeled as a path from the root to the leaf where each path from the root to the leaf represents one hand gesture.

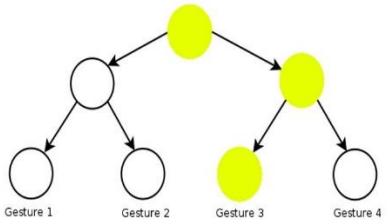


Fig.

Handling Redundant Gestures

The proposed tree based structure cannot generate a possible new gesture from the existing two entirely different hand gesture sequences represented by two trees having the possibility of a common sequence. Figure shows the example in which sequences 1,2,3,4 define one gesture and 4,5,6,7 define a second gesture, which depict two different poses. However since there is a possibility of sequence 4 arising after sequence 7 we can define a new sequence 5,6,7,4 thereby augmenting our training set.

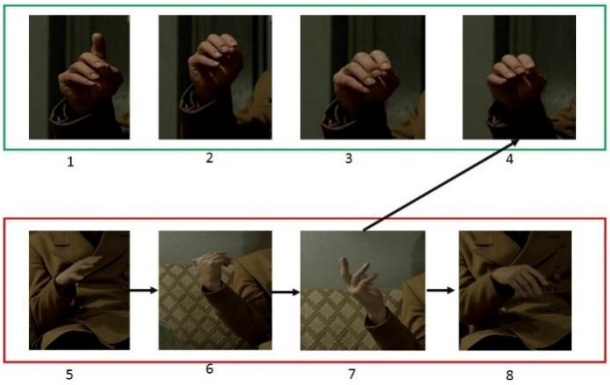


Fig.

To generalize the above phenomena, we calculate the similarity distance between nodes at level *i* of tree *j* to level *i*+1 of other trees of a forest within a gesture cluster. The most similar nodes are linked to reduce complexity in the forest structure with an assumption that intersecting frame always be an intermediate node of the gesture cluster, which is also proved through experiments. For each node *k* its corresponding

next probable gesture frame occurs more frequently at further levels of node *k*. On the other hand it is true that the next probable frame for a particular gesture node in other trees would appear with very less probability at the same level at which that current gesture is in the tree.

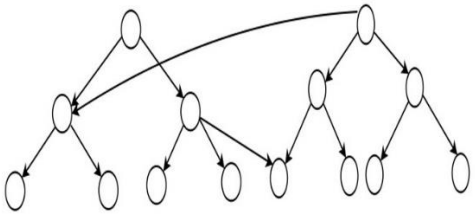


Fig.

Gesture Clustering

The second problem is to grouping massive amount of gestures into forests thereby creating a structure that would not be feasible for fast classification. To handle this problem, we use the concept of cluster of gestures using the training set effectively. We consider sets of hand sequences where each set consists of many trees, sequence of time frames denoting hand gestures, each starting from a common frame. That common frame is the identifier for the all hand gestures in a particular set. When two gestures within that set but be, there is a path from that node to another node in some other tree where the common node lies. Thus each forest contains many interconnected paths from one tree to another tree sharing common sequence frames between two different hand gestures.

Since there can be a lot of hand gestures which are almost same like waving left and waving left and right, we group them into clusters where each cluster represents a superset of a set of gestures. Therefore, we can have a set of clusters each of which contain a set of forests denoting a set of gestures. This paper introduces this very concept of a cluster of forests, which to the best of our knowledge is a first in recognizing hand gestures. To effectively cluster the gestures we use the concept of PHOG similarity.

For assigning a node into a particular cluster, we create an SVM model using the root nodes of each of the clusters of trees. When an input sequence is given, we take the first image from the sequence and calculate its membership value belonging to the clusters using the SVM model and the cluster with the highest membership value gets associated to the node is assigned to that node.

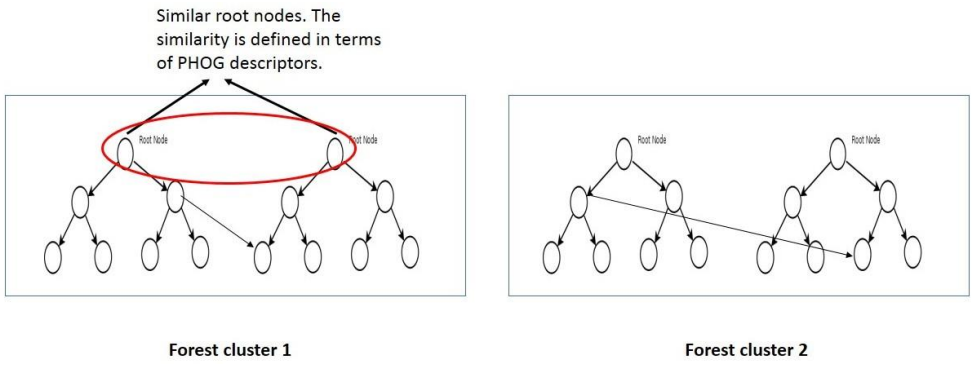


Fig.

Classification stage

The SVM model has been used in [] to classify the objects based on the HOG descriptor used by Dalal and Triggs. Even in PHOG, the SVM classifier is used where the kernel based on chi-square distance is used. We adopt a similar approach for classification except that this time we use the same distance measure as a threshold for finding the best path from the root to the leaf in a forest. The kernel measure is just a similarity measure to find the parent child distance measures between two consecutive frames of a gesture sequence.

For using the SVM model, we use the kernel that was implemented in [] based on pyramid levels. ed as

$$K(SI , SJ) = \alpha l \; dl \; (SI , SJ)$$

where αl is the weight at level l and dl the distance between SI and SJ at pyramid level l computed using $\chi 2$ (though other distances could be used). This defines the kernel for PHOG similarity. In the original spatial pyramid representation [14] each level was weighted using $\alpha l = 1/2(L-l)$ where L is the number of levels and l the current level. This means that histograms from finer resolutions are weighted more highly than those at coarser resolutions. However, this may not be the optimum weight choice for classification performance.

For our case the value of L is 3. We have used SVM in two cases. For the first phase, we first manually take 10 classes of hand gestures and manually map each of the training sequences to one gesture frame each.

We use the concept of multi-class classification for preparing the SVM model where a one-vs all strategy is used as identifiers for the gesture clusters. The reason behind using the gesture cluster is because of the fact that the millions of gestures that a person can exhibit to express his emotions or react to an environment makes it all the more difficult to interpret it properly if they are not clustered in the initial stages. But since we are avoiding any generic clustering algorithm, we shift to using SVM models for classifying each gesture into a pre-built initial set of gestures categorizing them into a very lean semantic model. The chi-square model and the RBF kernel both have been used extensively for the SVM model but keeping in line with that used in the spatial pyramid used in [], we use the chi-square kernel for classification based on the PHOG descriptors. So for each cluster we take the root node of all the trees within that gesture and tag it with a common label. We perform a similar operation for all other clusters as well. So the SVM training model consists of root nodes of all the trees but tagged with the label of their corresponding cluster to which they belong.

For classifying the test sequence into one of the gesture clusters, we take the first frame from the test sequence and perform SVM classification on the model that we prepared previously. Once the gesture cluster is identified, we create a second SVM model using the root nodes of the trees in that cluster in a similar manner we did for the clusters except that in this case we create a model using the root nodes of the trees in that cluster so as to achieve a starting point for traversal.

The prime purpose of having two SVM models is to cut down on the cost of traversing all the gestures in order to find the kind of gesture it resembles to by comparing it with all kinds of gestures available and instead limiting it to clusters.

Starting from the root node selected from the SVM model above, we follow these steps:

1. Queue q = root node.
    a. For node i = dequeue(q)
        i. For each child connected to node i, we calculate the chi-square distance between the current test frame and the node i.

ii. If the distance is lesser than the edge weight between the node i and child in consideration, then we enqueuethat child node in Queue q for further traversal and continue with rest of the child nodes of node i.

iii. Also map all child nodes to node i which satisfy the threshold condition in ii for retrieving the paths.

Case of missing frames:

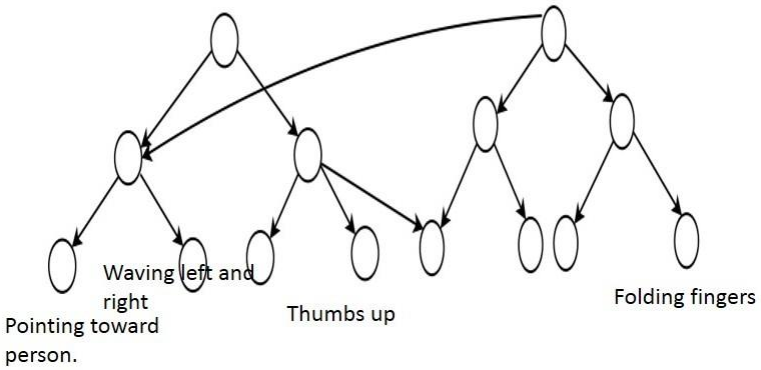
The one deficiency in a matching algorithm as ours using the forest datastructure is the case of missing frames or skipped frames in the test sequence. This could happen frequently while capturing frames in a video sequence as sometimes hand movements may be fast. To deal with this problem, we used the concept of string subsequence matching. We encode each path in our forest as a string

In case of missing frames, the following is performed:

1) If a particular gesture frame within a sequence cannot be matched with any of the nodes in the current level of the forest, then we consider all children of the next level of that tree and match that frame to those nodes of that level.
2) If we find a similarity below the threshold on that level, we continue with the next frame of the test sequence and the children below the matched node in the tree.
3) Else we continue with Step 1) but with the next level t+1 of the tree.
4) We continue this until any leaf node is reached.

In this way we would arrive at a string of nodes which must be a subsequence of one of all the possible strings within that forest. It is easy to find that the string extracted from the above step would be the subsequence of exactly one string from all the combinations within the tree. This is because every path from the root to the leaf would be encode exactly one string.

We annotate each path in the forest structure with one gesture alignment. The below figure shows this:



Each leaf is annotated with a gesture alignment that describes the sequence. Hence when we perform the same operation over the left hand and the righ hand we are left with the following information:

Left hand -> Gesture 1 and right hand-> Gesture 2.

We concatenate them so as to facilitate classification in the third phase, to produce "left hand gesture1 and right hand gesture 2". This text would be fed to the third phase to interpret social behavior.

Phase 3 – Augmenting the discrete semantic texts into behavioral actions

This is the third phase and it is this phase that augments the discrete semantic meanings obtained through the forest structure into proper behavioral actions which depicts the social behavior of the person communicated through the social signals based on hand gestures. The text mining approach to deciphering the meaning out of the attributes shown through the gestures involves latent semantic indexing, probabilistic lsa, Hidden Markov Models(HMM).

The main challenge in this kind of a problem is the task of preparing semantic lexicons that could properly capture the attributes involved in opinion mining from hand gestures. So in order to augment the hand alignment description from a pair of hands into a semantic interpretation of the person's social behavior, we need to co-align speech and gesture into a training corpus as has been described in []. We referred to the AMI corpus which contains recorded meetings and annotations of the meetings. We capture all the annotations from the corpus and prepare the bag of semantic lexicons described in the next sections.

Preparing the training model:

The table below shows the AMI corpus where column 1 and 2 are the speech and the corresponding alignment and column 3 depicts the high level social behavior of the person. We collect all the gesture descriptions from the corpus and tag it manually to 10 different social behaviors that are most commonly observed during communications.

Dialogue	**Gesture description**	**Social behavior**
Do you want to switch places	While C speaks, her right hand has its index finger extended; it starts at her waist and moves horizontally to the right towards D and then back again to C's waist, and this movement is repeated, as if to depict the motion of C moving to D's location and D moving to C's location.	Exclamatory situation
You walk out the doors	The gesture is one with a flat hand shape and vertical palm, with the fingers pointing right, and palm facing outward.	Firm order situation

For preparing an exhaustive set of sematic lexicons from the annotated gesture description in the corpus we perform the following:

1. tf –idf weighting: We used the traditional method of tf-idf weighting. Term Frequency - Inverse Document Frequency is a weighting scheme that is commonly used in information retrieval tasks. The goal is to model each document into a vector space, ignoring the exact ordering of the words in the document while retaining information about the occurrences of each word.

It is composed by two terms: one first computes the normalized Term Frequency, which is the number of times a word appears in a document, divided by the total number of words in that document. Then, the second term is the Inverse Document Frequency, which is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the term ti appears.

The TF-IDF gives how important is a word to a document in a collection, since it takes in consideration not only the isolated term but also the term within the document collection. The intuition is that a term that occurs frequently in many documents is not a good discriminator. So it scales down the frequent terms while scaling up the rare terms.

For computing the TF-IDF weights for each document in the corpus, we follow these steps: 1) Tokenize the corpus 2) Model the Vector Space and 3) Compute the TF-IDF weight for each document in the corpus.First we need to tokenize the text. First the text is splint into sentences, and then the sentences are split into the individual words. It is important to notice that there are several words that are not relevant, that is, terms like "the, is, at, on", etc... aren't going to help us, so in the information extraction, we ignore them. Those words are commonly called stop words and they are present in almost all documents, so it is not relevant for us.

Now that each of the documents in the corpus has been tokenized, the next step is to compute the document frequency quantity, that is, for each term, how many documents that term appears in. Before going to IDF, it is important to normalize the term-frequencies.

The TF-IDF is the product between the TF and IDF. So a high weight of the tf-idf is reached when there is a high term frequency (tf) in the given document and low document frequency of the term in the whole collection.

Using a threshold value, we can remove all such terms whose tf-idf value is less than that threshold.

2. **POS tagging:**Since sentiment words are often the dominating factor for sentiment classification, it is not hard to imagine that sentiment words and phrases may be used for sentiment classification in an unsupervised manner. We perform classification based on some fixed syntactic patterns that are likely to be used to express opinions.The syntactic patterns are composed based on part-of-speech (POS) tags.

The list of tags as used in [] are shown below:

CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign Word
IN	Preposition
SYM	Symbol
JJ	Adjective
MD	Modal
VBG	Verb
NN	Noun
RB	Adverb

The approach followed for extracting words after the tf-IDF approach is:

Step 1: Two consecutive words are extracted if their POS tags conform to any of the patterns in the table

1	JJ	NN/NNS
2	RB/RBR/RBS	JJ
3	JJ	JJ
4	RB/RB/RBS	VB/VBD/VBN
5	NN	JJ

For example, pattern 2 means that two consecutive words are extracted if the first word is an adverb, the second word is an adjective. As an example, in the sentence “He is pointing toward the writing board”,

“pointing toward” is extracted as it satisfies pattern 5. The reason these patterns are used is that JJ, RB, RBR and RBS words often express opinions. The nouns or verbs act as the contexts because in different contexts a JJ, RB, RBR and RBS word may express different sentiments.

For example, the adjective (JJ) “unpredictable” may have a negative sentiment in a car review as in “unpredictable steering,” but it could have a positive sentiment in a movie review as in “unpredictable plot.”

Using the above method we build a bag of sentiment lexicons from a large number of annotated text documents such that each social behavior is mapped to a bag.

3. **Classification:** In order to assign a social behavior to the discrete gesture alignment obtained from the tree structure, we use the concept of sentiment orientation measured through PMI (Pointwise Mutual Information).

PMI measures the degree of statistical dependence between two terms.Here, Pr(term1, term2) is the actual co-occurrence probability of term1 and term2. Hence we find the PMI between each bigram of the input phrase and the bag of semantic lexicons corresponding to each behavior. We then find the maximum among them and classify that input to display that particular behavior.