# CHAPTER 2

# LITERATURE SURVEY

[1**] Low Rank Approximation**

This paper proposes a framework based on the Hidden Markov Models (HMMs) benefited from the low rank approximation of the original sign videos for two aspects. First, under the observations that most visual information of a sign sequence typically concentrates on limited key frames, we apply an online low rank approximation of sign videos for the first time to select the key frames. Second, rather than fixing the number of hidden states for large vocabulary of variant signs, we further take the advantage of the low rank approximation to independently determine it for each sign to optimise predictions. With the key frame selection and the variant number of hidden states determination, an advanced framework based on HMMs for Sign Language Recognition (SLR) is proposed, which is denoted as Light-HMMs (because of the fewer frames and proper estimated hidden states). With the Kinect sensor, RGB-D data is fully investigated for the feature representation. In each frame, we adopt Skeleton Pair feature to character the motion and extract the Histograms of Oriented Gradients as the feature of the hand posture appearance. The proposed framework achieves an efficient computing and even better correct rate in classification. The widely experiments are conducted on large vocabulary sign datasets with up to 1000 classes of signs and the encouraging results are obtained.

**[2] Sign Language Recognition System**

In this paper A gesture may be defined as a movement, usually of hand or face that expresses an idea, sentiment or emotion e.g., rising of eyebrows, shrugging of shoulders are some of the gestures we use in our day-to-day life. Sign language is a more organized and defined way of communication in which every word or alphabet is assigned some gesture. In American Sign Language (ASL) each alphabet of English vocabulary, A-Z, is assigned a unique gesture. Sign language is mostly used by the deaf, dumb or people with any other kind of disabilities. With the rapid advancements in technology, the use of computers in our daily life has increased manifolds. Our aim is to design a Human

Computer Interface (HCI) system that can understand the sign language accurately so that the signing people may communicate with the non-signing people without the need of an interpreter. It can be used to generate speech or text. Unfortunately, there has not been any system with these capabilities so far. A huge population in India alone is of the deaf and dumb. It is our social responsibility to make this community more independent in life so that they can also be a part of this growing technology world. In this work a sample sign language has been used for the purpose of testing. No one form of sign language is universal as it varies from region to region and country to country and a single gesture can carry a different meaning in a different part of the world. Various available sign languages are American Sign Language (ASL), British Sign Language (BSL), Turkish Sign Language (TSL), Indian Sign Language (ISL) and many more. There are a total of 26 alphabets in the English vocabulary. Each alphabet may be assigned a unique gesture. In our project, the image of the hand is captured using a simple web camera. The acquired image is then processed and some features are extracted. These features are then used as input to a classification algorithm for recognition. The recognized gesture may then be used to generate speech or text. Few attempts have been made in the past to recognize the gestures made using hands but with limitations of recognition rate and time. This project aims at designing a fully functional system with significant improvement from the past works.

**[3] Continuous Sign Language Recognition**

This paper introduces the end-to-end embedding of a CNN into a HMM, while interpreting the outputs of the CNN in a Bayesian fashion. Gesture is a key part in human to human communication. However, the role of visual cues in spoken language is not well defined. Sign language on the other hand provides a clear framework with a defined inventory and grammatical rules that govern joint expression by hand (movement, shape, orientation, place of articulation) and by face (eye gaze, eye brows, mouth, head orientation). This makes sign languages, the natural languages of the deaf, a perfect test bed for computer vision and human language modelling algorithms targeting human computer interaction and gesture recognition. Videos represent time series of changing images. The recognition of sign language therefore needs to be able to cope with variable input sequences and execution speed. Different schemes are followed to achieve this ranging from sliding window approaches to temporal normalisations or dynamic time warping. While in the field of automatic speech recognition, HMM dominate the field, they remainx rather unpopular in computer vision related tasks. This may be related to the comparatively poor image modelling capabilities of Gaussian Mixture Models (GMMs), which are traditionally used to model the observation probabilities within such a framework. More recently, deep Convolutional Neural Networks (CNNs) have outperformed other approaches in all computer vision tasks. Which is why we focus on integrating CNNs in a Hidden-Markov-Model (HMM) framework, extending an interesting line of work. For the field to have more realistic scenarios, such as those imposed by challenging real-life sign language corpora, are required. We presented a hybrid CNN-HMM frame-work that combines the strong discriminative abilities of CNNs with the sequence modelling capabilities of HMMs, while abiding to Bayesian principles.

**[4] Real Time Hand Gesture Recognition**

The scale invariant feature transform (SIFT) algorithm, developed by Lowe [1,3,4], is an algorithm for image features generation which are invariant to image translation, scaling, rotation and partially invariant to illumination changes and affine projection. SIFT algorithm can be used to detect distinct features in an image. Once features have been detected for two different images, one can use these features to answer questions like “are the two images taken of the same object?” and “given an object in the first image, is it present in the second image?”. Computation of SIFT image features is performed through the four consecutive phases which are briefly described in the following. In order to get a reliable recognition, it is quite important that the features extracted from the training image are detectable even under changes in image scale, noise and illumination. Such points generally lie on high-contrast regions of the image, for example object edges. Gesture recognition is initially performed by matching each keypoint independently to the database of key points extracted from training images. Many of these initial matches will be incorrect due to ambiguous features or features that arise from background clutter. Therefore, clusters of some features are first identified that agree on an object and its pose, as these clusters have a much higher probability of being correct than individual feature matches. Then, each cluster is checked by performing a detailed geometric fit to the model, and the result is used to accept or reject the interpretation.