Human Action Recognition in Video by 'Meaningful' Poses

Snehasis Mukherjee*
Electronics and Communication Sciences Unit,
Indian Statistical Institute
203 B.T. Road
Kolkata-700108, India
snehasismukho@gmail.com

Sujoy Kumar Biswas
Electronics and Communication Sciences Unit,
Indian Statistical Institute
203 B.T. Road
Kolkata-700108, India
skbhere@gmail.com

Dipti Prasad Mukherjee
Electronics and Communication Sciences Unit,
Indian Statistical Institute
203 B.T. Road
Kolkata-700108, India
dipti@isical.ac.in

ABSTRACT

We propose a graph theoretic technique for recognizing actions at a distance by modeling the visual senses associated with human poses. Identifying the intended meaning of poses is a challenging task because of their variability and such variations in poses lead to visual sense ambiguity. Our methodology follows a bag-of-words approach. Here "word" refers to the pose descriptor of the human figure corresponding to a single video frame and a "document" corresponds to the entire video of a particular action. From a large vocabulary of poses we prune out ambiguous poses and extract 'meaningful' [6] poses - for each action type in a supervised fashion - using centrality measure of graph connectivity [16]. The number of 'meaningful' poses per action is determined by setting a bound on the centrality measure. We evaluate our methodology on four standard activity recognition datasets and the results clearly demonstrate the superiority of our approach over the present state-of-the-art.

1. INTRODUCTION

Human action recognition in image and video is an active area of research. The initiatives in this field usually have two broad classifications - either they focus on low and mid-level feature collection ([10, 17]) or they model the high level interaction among the features [14, 13]. For example, Mori *et. al.* have proposed a learned geometric model to represent human body parts in an image, where the action is recognized by matching the static postures in the image with the target action [14, 13]. Similarly, Cheung *et. al.* have used the silhouette of the body parts to represent the shape of the performer [3]. Recently, the bag-of-words model is being used to recognize actions in videos [10, 17]. Shah *et. al.* have used a vocabulary of local spatio-temporal volumes (called cuboids) and a vocabulary of spin-images (to capture the shape deformation of the actor by considering actions as 3D objects (x, y, t)) [10]. Niebles *et. al.* also uses some space-time interest points on the video as features (visual words) [17]. The algorithm of Niebles *et. al.*, automatically learns the probability distributions of the visual words using graphical models like probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) to form the vocabulary of words. In [20], the whole frame in a video is represented as "word", instead of "collection of words" as proposed in [17] in the bag-of-words model. The main success in the work of Mori *et. al.* is that, they have successfully applied the method on some data where the object (the actor) is very small (30 to 40 pixels).

Bag-of-word based action recognition tasks either seek right kind of features for video words or model the abstraction behind the video words. There are initiatives which study pose specific features [8] but modeling visual senses associated with poses in videos is largely an unexplored research area. The proposed methodology is built here on the following premise - human poses often carry a strong visual sense (intended meaning) which describes the related action unambiguously. But this is a challenging task given the tremendous variation present in the visual poses either in form of external variation (viz. noise, especially in low resolution videos) or variation inherent to the human poses. Variation in poses is the primary source of visual sense ambiguity and often a single pose gives out a confusing interpretation about the related action type. For example, top row in Figure 1 shows some ambiguous poses and by looking at them one cannot tell for certain the corresponding actions, whereas the bottom row illustrates the 'meaningful' poses which unambiguously specify the related actions. Our contribution in this paper is two-fold. First, we propose a novel pose descriptor which not only captures the pose specific details of the performer from a single video frame but also considers motion information of the poses from two subsequent frames. Secondly, we seek to model visual senses exhibited by the human poses. For each visual sense (i.e., action type) we rank the poses in order of "importance" using centrality measure of graph connectivity [16]. The emphasis on the pose specific details is in accordance with the theme of this paper - action recognition at a distance; we argue that when the camera is placed far from the performer it becomes difficult to model each part of his/her body separately. The foreground figure of the human performer appears as a tiny blob (of approximate height of 40 pixels) and the only reliable cue offered in such low-resolution videos is given by the motion pattern of the poses.

The proposed methodology consists of combining the motion and pose information of a human performer into a single

*Corresponding author

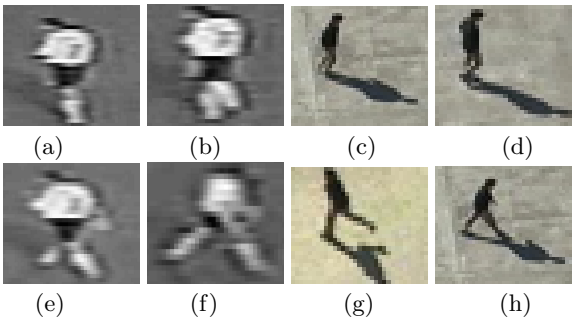

Figure 1: Top row shows some ambiguous poses (as labeled by our algorithm) from (a), (b) Soccer dataset and (c), (d) Tower dataset. The bottom row shows retrieved 'meaningful' poses (by our algorithm) for (e), (f) walking from Soccer dataset, (g) running and (h) walking actions from Tower dataset

multi-dimensional descriptor. This is done by deriving local histograms of oriented flow vectors from a weighted optical flow field and then concatenating the local histograms into a single global pose descriptor. The global pose descriptor corresponds to a single video frame. The pose descriptors are full with redundancies and upon clustering we obtain an over-complete codebook of visual poses. The pose clusters may be equated with visual words and documents here stand for an entire video sequence. A sparse set of discriminatory poses are selected from this over-complete codebook in a supervised fashion, i.e. this set of poses is constructed separately for each action type starting from the over-complete codebook. Such discriminative set of poses obtained by eliminating ambiguous poses from the over-complete codebook is called compact codebook. This sparse set of poses are used for classification of an unknown target video.

The sparse set of poses are obtained from the over-complete dictionary by the feature ranking technique known as the centrality measure of graph theory. This requires the construction of a pose graph corresponding to each action type where the pose graph contains poses from the over-complete codebook as vertices and an edge between two vertices explains the joint behavior of the two poses. By joint behavior we mean how well the two poses describe the action together. Next we rank the poses using the graph centrality measure and then choose the most 'important' or 'meaningful' poses for a particular kind of action using the concept of 'meaningfulness' [6]. Grouping all such poses together, we build our sparse codebook.

Section 2 presents the proposed methodology. The results in Section 3 show the efficiency of the proposed approach. In Section 4, we draw our conclusions.

2. PROPOSED APPROACH

As discussed in the Introduction, our first task is to derive a multidimensional vector (called the pose descriptor) corresponding to each frame of all the video. The pose descriptors, upon data condensation result into a moderately compact representation and we call it an over-complete codebook of visual poses. From the over-complete codebook, a relatively compact set of visual words is formed by selecting only some 'meaningful' poses which can uniquely identify a particular action. We get a histogram corresponding to each

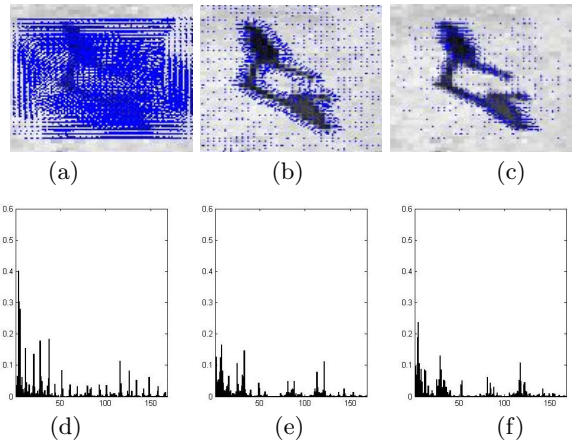

Figure 2: (a) The optical flow field, (b) gradient field, (c) weighted optical flow and (d), (e), (f) show the respective pose descriptor (histograms obtained from (1)) on a frame of a sample video.

action video, showing the frequency of each 'meaningful' poses in the video. We learn the histograms corresponding to all the action videos and test the query video by matching the histograms. So first we discuss the methodology for deriving the descriptors.

2.1 Deriving the Pose Descriptor

Our pose descriptor combines the benefit of motion information from optical flow field (using Lucas-Kanade algorithm [12]) and pose information from the gradient field. We produce a flow field vector $\vec{V}$ from the optical flow field $\vec{F}$, weighted with the strength of the gradient field $\vec{B}$, i.e.,

$$\vec{\mathbf{V}} = |\vec{\mathbf{B}}|. * \vec{\mathbf{F}}, \tag{1}$$

where the symbol '.*' represents the point wise multiplication of the two matrices.

The effect of this weighted optical flow field is best understood if one treats the gradient field as a band pass filter. This is because the gradient field takes high value where the edge is prominent, preferably along the boundary of the foreground object, but it is very low in magnitude on the uniform background space. Since gradient strength along the human silhouette is quite high, the optical flow vectors there get a boost upon modulation with gradient field strength. So we filter in the motion information along the silhouette of the human figure and suppress the flow vectors elsewhere in the frame. So our descriptor is basically a motion-pose descriptor preserving the motion pattern of the human pose. Figure 2 shows how our descriptor differs from the optical flow field and the gradient field and gives more importance to the movement of the human silhouette and minimizes the effect at the other points in the frame.

Suppose we have a video sequence of some action type A having M frames and we denote the frames of the sequence by $I_1, I_2, ..., I_M$. The frame $I_i, i \in 1, 2, ..., M$ is a grey image matrix defined as a function such that for any pixel (x, y), $I(x, y) \in \Theta$, where $(x, y) \in Z^2$ and $\Theta \subset Z^+$ determines the range of the intensity values. Corresponding to each pair of consecutive frames I_{i-1} and I_i, $i \in 1, 2, ..., M$, we compute the optical flow field $\vec{F}$. Also we derive the gradient

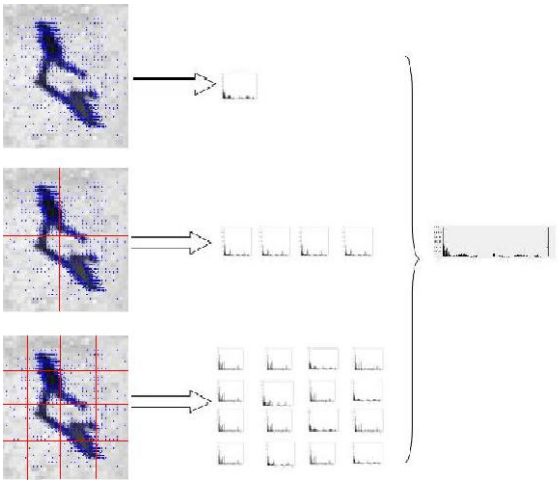

Figure 3: Formation of 168-dimensional pose descriptor in three layers.

field vector $\vec{B}$ corresponding to frame I_i and following (1) we obtain $\vec{V}$. We consider a three layer image pyramid (Figure 3), where in the topmost layer we distribute the field vectors of $\vec{V}$ in an L-bin histogram. Here each bin denotes a particular octant in the angular radian space. We take the value of L as 8, because orientation field is quantized enough when resolved in eight directions, i.e., in every 45 degrees. The derived field $\vec{V}$ can be resolved in two channels V_x and V_y along the x and y-components respectively, i.e. $\vec{V} = (V_x, V_y)$. The histogram $H = \{h(1), h(2), ..., h(L)\}$ construction takes place by quantizing $\theta(x, y) = \arctan(\frac{V_y}{V_x})$ and then adding up $m(x, y) = \sqrt{V_x^2 + V_y^2}$ to the right bin indicated by the quantized θ. In mathematical notation the process is as follows.

$$h(i) = \sum_{x,y} \begin{cases} m(x,y) & when\ \theta \in i\text{th octant} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The next layer in the image pyramid splits the image into 4 equal blocks and each block produces one 8-bin histogram leading to 32-dimensional histogram vector. Similarly, the bottommost layer has 16 blocks and hence produce 128-dimensional histogram vector. All the histogram vectors are *L1*-normalized separately for each layer and concatenated together resulting in a 168-dimensional pose descriptor. Once we have the pose descriptors we seek to quantize them into a visual codebook of poses. Next section outlines the details of the visual codebook formation process.

2.2 Formation of Visual Codebook

Since human action has repetitive nature, an efficient pose descriptor derived above retains redundancy. Instead of clustering, we solicit the idea of data condensation because in data condensation one may afford to select multiple prototypes from the same cluster, whereas in case of clustering one seeks to identify the true number of clusters (and also their true partitioning). The data condensation ultimately leaves us with an over-complete codebook S of visual poses where at least some of the redundancies of the pose space is eliminated. The learning of the pose codebook follows Maxdiff *kd*-tree based data condensation technique [15].

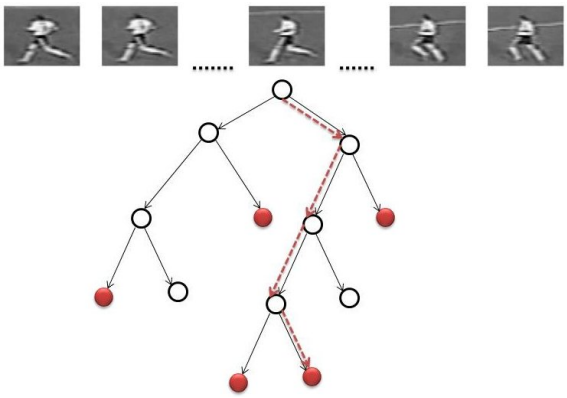

Figure 4: Mapping of a pose descriptor to a pose word in the *kd*-tree; leaf nodes in the tree denote poses and red leaf nodes denote 'meaningful' poses.

An optimum (local) lower bound on the codebook size of S can be estimated by Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) [1] or one can directly employ X-means algorithm [18] which is a divisive clustering technique where splitting decision depends on the local BIC score. X-means based clustering techniques [18] rely on Euclidean distance metric which is isotropic in nature and perform poorly when the dimension of the feature space increases [1]. The Maxdiff *kd*-tree based data condensation technique alleviates the curse of dimensionality by mining the multi-dimensional pose descriptors into a *kd*-tree data structure. The leaf nodes of the *kd*-tree denote one pose cluster or the visual pose word; one can choose (depending on computational expense) multiple samples from each leaf node to construct the large pose vocabulary $S = \{p_i \in \Re^d \mid i = 1, 2, ..., k\}$, where d denotes the dimensionality of the pose descriptors and k is the cardinality of S. The algorithm to construct the *kd*-tree is explained in details in [15]. In our experiment we choose the mean of each leaf as our pose word and learn the codebook S of poses. The pose descriptor in a video sequence is mapped to a pose word in S by descending down the *kd*-tree (by the tree traversal algorithm) and hitting the leaf node (Figure 4). If one selects multiple poses from the same leaf node in the construction of S, one can break the tie (in word - descriptor mapping) by computing the nearest neighbor of the pose descriptor. Next we outline the scheme for ranking the poses of S using centrality theory of graph connectivity.

2.3 Pose Ranking by Centrality Measure of Graph Connectivity

The poses in the over-complete visual codebook S are often ambiguous and our goal is to identify the unambiguous (i.e., 'meaningful') poses. The poses from S are embedded in a graph as nodes and the edge between each two poses stands for the dissimilarity in terms of a semantic relationship between them, measured using some form of weight function. The idea of "importance" of a pose is represented based on the notion of centrality - a pose is central if it is maximally connected to all other poses. The 'meaningful' poses are identified in the graph separately for each action repeating the same algorithmic procedure for all kinds of ac-

tions. We define a pose graph for a specific kind of actions as follows:

Definition 1. A pose graph for an action is an undirected edge-labeled graph $G = (S, E)$ where each vertex in G corresponds to a pose belonging to the over-complete codebook S; E is the set of edges and $\omega : E \rightarrow (0, 1]$ is the edge weight function. There is an undirected edge between the poses u and v ($u \neq v$ and $u, v \in S$), with edge weight $\omega(u, v)$, iff $0 < \omega(u, v) \leq 1$. Edge weights indicate dissimilarity between the two poses. It is assumed that ω is symmetric i.e., $\omega(u, v) = \omega(v, u)$, $\forall\ u, v \in S$ and $\omega(u, u) = 0\ \forall\ u \in S$.

As discussed earlier, human activity follows a sequence of pose patterns in definite order and they have a cyclic nature. For simplicity we have assumed the cycle length fixed and used a span of T frames to define an action cycle. Most of the repetitive actions in our datasets (like running, walking, jumping, etc.) complete a full cycle in and around 10 frames and we set the value of T at 10. Let $\rho(u, v)$ denote how many times the pose words u and v both occur together in all the action cycles of a particular action video.

$$\omega(u, v) = \begin{cases} \frac{1}{\rho} & when\ \rho(u, v) \neq 0 \\ C & \text{otherwise} \end{cases} \tag{3}$$

where, C is a large constant when u and v do not have an edge in between them. According to (3), lower is the edge weight, stronger is the semantic relationship between the pose words. One can think it this way - same context causes these two poses happen together. Next we define the *eccentricity* measure [21] of graph connectivity to see how semantically different the poses are from each other in a pose graph.

Definition 2. Given a pose graph G, the distance $d(u, v)$ between two pose words u and v (where $u, v \in S$) is the sum of the edge weights on a shortest path from u to v in G. Eccentricity $e(u)$ of any $u \in S$ is the maximum distance from u to any $v \in S$; ($u \neq v$), i.e.,

$$e(u) = max\{d(u, v) \mid v \in S\}.$$

Floyd-Warshall algorithm [4] computes all-pair-shortest path to evaluate the *eccentricity* $e(u)$ of each pose$u \in S$ using Definition 2. According to Definition 2, *eccentricity* $e(u)$ is a measure of ambiguity of the pose u. So for each action, we choose the unambiguous poses by selecting poses with significantly low *eccentricity* value in a pose graph. The following proposition narrates the pose ranking based on *eccentricity* value:

Proposition 1: Given a graph connectivity measure "e" and the set of vertices S, for a pair of vertices $u, v \in S$, we induce a ranking $rank_e$ of u and v such that $rank_e(u) \leq rank_e(v)$ iff $e(u) \geq e(v)$.

Now the problem is, how many poses should be selected for compact dictionary for each action? One procedure may be to select q-best poses (in terms of lowest *eccentricity*) from each action to form the compact codebook. Then vary the number q in all integral values of some interval $[a, b]$ ($a, b \in Z^+$), to calculate the accuracy for each q. Lastly take the optimal q. The problem in this procedure is that, in reality, the number of unambiguous poses is usually different for each action type. The concept of meaningfulness [6] gives us the opportunity to vary the number of selected poses for different action type. We illustrate the process forming compact codebook ξ from the over-complete codebook S in the next subsection.

2.4 Formation of Compact Codebook Selecting 'Meaningful' Poses

Now we have a 168-dimensional vector (pose descriptor) corresponding to each frame of each action video. We also have an over-complete codebook of a number of pose words, each having an *eccentricity* value depicting a measure of ambiguity. We first normalize the *eccentricity* values between 0 and 1 by dividing all *eccentricity* values with their maximum. The poses with 'meaningfully' low *eccentricity* values are selected from the over-complete codebook to produce the compact codebook following the definition:

Definition 3. Given a sequence of mutually exclusive sets of action types $\{A_n\}_{n=1,2,...,\alpha}$ (A_n is the set of all pose descriptors of the nth action type, α is the number of action types) and an over-complete codebook S ($\bigcup_{n=1}^{\alpha}\{A_n\} = S$), a set $\xi \subset S$ is said to be a compact codebook if

$$(i)\ \xi = \{u \in S \mid e(u) < \delta\}\ and$$

$$(ii)\ \forall A_n,\ \exists\ u \in A_n\ such\ that\ u \in \xi,$$

where $e(u)$ is the *eccentricity* value of the word u and δ is a 'meaningful' cut-off value of $e(u)$.

2.4.1 Selecting Meaningful cut-off for eccentricity

This 'meaningful' cut-off value is determined by the concept of 'meaningfulness' introduced by Agnes *et. al.* [6]. The concept of meaningfulness is derived from the Gestalt Philosophy. The Gestalt hypothesis is being used to solve several problems in the field of computer vision [6]. According to the Gestalt theory, "grouping" is the main concept for our visual perception [5]. Suppose there are n objects, k of them having similar characteristics with respect to some *a priori* knowledge (e.g., same color, same alignment, etc.). Then the question is that, are these characteristics happening by chance, or is there any significant cause to group them to form a meaningful characteristics? To answer this question, we first assume that the characteristics are uniformly distributed over all the n objects and the observed characteristics are some random realization of the uniform distribution. According to the concept of meaningfulness, if the expectation of the observed configuration of k objects is very small, then the grouping of these objects is meaningful. We calculate the expected number of occurrences of the observed characteristics, which is called the number of false alarms (NFA). If the NFA is less than a certain number ϵ then observed characteristic is called an ϵ-meaningful event; otherwise it is a random event. That is by definition, a meaningful event is significantly different from random events and has a very small NFA. In this paper, our object is to select from each action type, only some poses having 'meaningfully' low $e(u)$ value. For simplicity in further calculation, we introduce another measure $E(u) = 1 - e(u)$. So now our object is to find a cut-off $\Delta = 1 - \delta$, hence the poses having 'meaningfully' high $E(u)$ value (greater than Δ) are selected as 'meaningful' poses.

For finding the meaningful cut-off value η of the *eccentricity* value $E(u)$ for each action type A_n ($n \in \{1, 2, ..., \alpha\}$), we

first select for each A_n, λ equidistant points in [0,1], the range of $E(u)$ value. We vary the value of the threshold η over all the chosen equidistant points in [0,1]. For each η we do the following two steps given by the two equations (4) and (5). If ν is the prior probability that an arbitrary pose has $E(u)$ higher than η, then

$$\nu = 1 - \eta, \tag{4}$$

assuming that the values of $E(u)$ are i.i.d. uniformly distributed in the interval [0,1].

Let t be the minimum number of poses needed to recognize the action type A_n [5]. The cut-off η is meaningful if the action type A_n contains at least t poses due to cut-off η. Therefore, whether a particular cut-off value is meaningful or not becomes a Bernoulli trial. If $(1-P_n)$ is the probability that the cut-off η is meaningful, then

$$P_n^\eta = \sum_{i=t}^{M} \binom{M}{i} \nu^i (1-\nu)^{M-i}, \tag{5}$$

is the Binomial tail, where M is the number of poses in A_n and ν comes from (4).

Here the problem is a Bernoulli trial (in our problem, whether a cut-off value of $E(u)$ is meaningful or not), then the NFA of the event that the particular cut-off η is significant for detecting 'meaningful' poses, can be defined as

$$NFA = \lambda P_n^\eta, \tag{6}$$

where P_n^η comes from (5). λ is the number of equi-spaced η values in the interval (0,1) to estimate the meaningful η. In other words, λ is the number of trials. If the value of NFA is less than a predefined number ϵ, then the corresponding cut-off value η is ϵ-meaningful. Setting $\epsilon = 1$ as in [6], means that the expected number of occurrence of the event that, the cut-off $\eta = \eta'$ is meaningful for the corresponding action type, is less than 1. Let us call this cut-off η' as '1-meaningful' cut-off value.

2.4.2 Estimation of Parameters for Finding Meaningful Cut-off

Parameter λ is used to calculate the NFA in (6) and P_n^η is obtained by (5). Then the probability that all the poses in the action type A_n have $E(u)$ greater than η, is ν^M. This is lesser or equal to the probability that at least t poses have $E(u)$ less than η, which is P_n^η. So $\nu^M \leq P_n^\eta < \frac{\epsilon}{\lambda}$ (since according to (6), for an ϵ-meaningful event, $NFA = \lambda P_n^\eta < \epsilon$), which implies

$$M \geq \frac{\log \epsilon - \log \lambda}{\log \nu}. \tag{7}$$

For a given η, ν comes from (4). M is fixed for a given action type A_n. Then for a given ϵ, we can find λ from (7).

Parameter t is the minimum number of poses needed to recognize the action type A_n. From Hoeffding's inequality [9], for an ϵ-meaningful event we can deduce the following:

$$t \geq \nu M + \sqrt{\frac{M}{2}(\log \lambda - \log \epsilon)}. \tag{8}$$

The equation (8) is the sufficient condition of meaningfulness. The derivation comes from Hoeffding's inequality.

PROOF. In our problem, M is the number of poses in action type A_n in the codebook. We can formulate the problem by an i.i.d. sequence of M random variables $\{X_q\}_{q=1,2,3,\ldots,M}$, such that $0 \leq X_q \leq 1$. Let us define X_q as,

$$X_q = \begin{cases} 1 & when\ E(q) < \eta \\ 0 & \text{otherwise} \end{cases}$$

for a given η. We set $S_M = \sum_{q=1}^{M} X_q$ (i.e., the number of poses of A_n having $E(u)$ value greater than η) and $\nu M = E[S_M]$. Then for $\nu M < t < M$ (since ν is a probability value less than 1), putting $\sigma = \frac{t}{M}$ as in [6], according to Hoeffding's inequality,

$$P_n^\eta = P(S_M \geq t) \leq e^{-M(\sigma \log \frac{\sigma}{\nu} + (1-\sigma)\log \frac{1-\sigma}{1-\nu})}.$$

In addition, the right hand term of this inequality satisfies,

$$e^{-M(\sigma \log \frac{\sigma}{\nu} + (1-\sigma)\log \frac{1-\sigma}{1-\nu})} \leq e^{-M(\sigma-\nu)^2 H(\nu)} \leq e^{-2M(\sigma-\nu)^2},$$

where

$$H(\nu) = \begin{cases} \frac{1}{1-2\nu} \log \frac{1-\nu}{\nu} & when\ 0 < \nu < \frac{1}{2} \\ \frac{1}{2\nu(1-\nu)} & \frac{1}{2} \leq \nu < 1 \end{cases}$$

This is Hoeffding's inequality. We then apply this for finding the sufficient condition of ϵ-meaningfulness. If $t \geq \nu M + \sqrt{\frac{\log \lambda - \log \epsilon}{H(\nu)}}\sqrt{M}$, then putting $\sigma = \frac{t}{M}$ we get

$$M(\sigma - \nu)^2 \geq \frac{\log \lambda - \log \epsilon}{H(\nu)}.$$

Then,

$$P_n^\eta \leq e^{-M(\sigma-\nu)^2 H(\nu)} \leq e^{-\log \lambda + \log \epsilon} = \frac{\epsilon}{\lambda}.$$

This means by definition of meaningfulness, that the cut-off η is meaningful (according to (6)).

Since for ν in (0,1), $H(\nu) \geq 2$ we get the sufficient condition of meaningfulness as (8). □

It is clear from (4) and (5) that if η' is a 1-meaningful cut-off value for $E(u)$ value, then each cut-off value chosen from the interval $[\eta', 1]$ is also 1-meaningful. Now from the 1-meaningful cut-off values, we have to select the maximal meaningful cut-off.

2.4.3 Selecting Maximal Meaningful Cut-off

Setting $\epsilon = 1$ is a safe choice to find the maximal meaningful cut-off. However, for choosing the maximal meaningful cut-off, we should have some measure of meaningfulness. For this purpose, we should consider the empirical probability of a pose of A_n to have its $E(u)$ to fall in the interval $[\eta', 1]$. Let $r_n(\eta')$ be the empirical probability of a pose of A_n to have $E(u)$ in the interval $[\eta', 1]$. Then

$$r_n(\eta') = \frac{M(\eta')}{M}, \tag{9}$$

where $M(\eta')$ is the number of poses in A_n having $E(u)$ greater than η'.

In general, for 1-meaningful cut-off values, $r_n(\eta') < \nu$. Now using $r_n(\eta')$, we have to define a measurement of maximal meaningfulness of the cut-off value. This measurement should penalize the situation that a 1-meaningful cut-off ζ yields higher empirical probability value than ν. This measurement (let us call it as c-value) should also help to reduce the number (not compromising with the accuracy of recognizing the action type) of 'meaningful' poses for an action

type. However, according to Definition 3, the corresponding action type must have at least one selected pose. Then c-value can be defined as,

$$c_n(\zeta) = \begin{cases} \infty & \\ & when\ r_n(\zeta) \geq \nu, \quad \text{or}, \quad r_n(\zeta) = 0 \\ r_n(\zeta) \log \frac{r_n(\zeta)}{\nu} + (1 - r_n(\zeta)) \log \frac{(1-r_n(\zeta))}{(1-\nu)} & \\ & \text{otherwise} \end{cases} \tag{10}$$

where ζ can take any value from the interval (η',1). We take the open interval instead of the closed interval, in order to avoid division by zero in (10), which occurs if $\nu = 0$.

For each A_n, we find the $c_n(\zeta)$ for all $\frac{1}{\lambda}$ distant values of ζ from the interval (η',1). Clearly, a more meaningful value of η' gives lesser value of $c_n(\zeta)$. For each action type, we find the maximal meaningful cut-off using the following definition:

Definition 4. A cut-off ζ is said to be maximal meaningful cut-off for the corresponding action type, if it is 1-meaningful and

$$\forall\ m \in (\eta', 1) - \{\zeta\}, \quad c_n(\zeta) \leq c_n(m).$$

The frames with $E(u)$ value greater than the maximal meaningful cut-off value of the corresponding action type, are finally chosen as 'meaningful' poses and included in the compact codebook ξ. Next we illustrate the results of our approach.

3. RESULTS AND DISCUSSIONS

The choice of dataset is made keeping in mind the focus of our paper - recognizing action at a distance. Soccer [7], Tower [2], Hockey [11] datasets contain human performer far away from the camera and 30-40 pixels tall approximately. Only exception is the KTH [19] dataset where we evaluated our proposed methodology on medium size (100 pixel tall) human figure. We use support vector machines for classification of target video with radial basis function following a "Leave-one-out" scheme, i.e., our training set consists of all of the action video sequences except the one which we hold out for evaluating our trained models. And we repeat this step for all the given video sequences. This same process is carried out for all the four datasets. In [7, 20] an automatic preprocessing step is used to centralize the human figure. There is no such requirement in our algorithm; the weighed optical flow vectors obtained by (1), get automatically magnified around the silhouette of the foreground figure (due to higher gradient strength) and subdued elsewhere because of lower gradient strength. For each dataset we show the confusion matrix of different codebook models.

Our approach is efficient both in terms of consumed time and accuracy in detecting human actions in comparison to state-of-the-arts (Table 1). The major time consuming step is learning the 'meaningful' poses for each action separately. But this is done once and we reap benefit later while classifying video with a small set of just 4 or 5 selected poses per actions. The average time consumed for learning 'meaningful' poses by each action amounts to little less than one minute. After the detection of 'meaningful' poses in S, our approach takes only a few seconds for both learning and testing in our MATLAB7 implementation in a machine with processor speed 2.37 GHz, 512MB RAM.

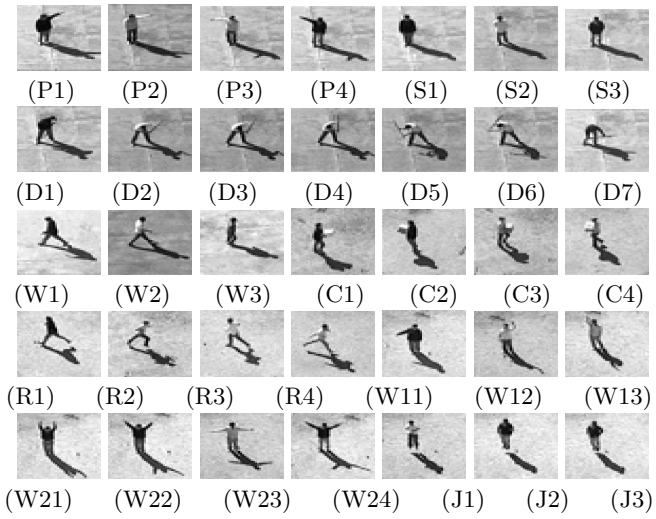

Figure 5: Selected 'meaningful' poses of the Tower dataset. (P1-P4) Pointing, (S1-S3) Standing, (D1-D7) Digging, (W1-W3) Walking, (C1-C4) Carrying, (R1-R4) Running, (W11-W13) Waving 1, (W21-W24) Waving 2, (J1-J3) Jumping.

Table 2 shows us how the concept of meaningfulness enhanced the efficiency of the proposed approach. This gives much better result than the result obtained by selecting some fixed number of poses for each action type to construct the compact codebook. Figure 5 shows the 'meaningful' poses for all the 9 action types of the Tower dataset.

3.1 Soccer dataset.

The Soccer dataset contains several video sequences of digitized World Cup football game from an NTSC video tape [7]. In this dataset each video sequence has more than one action. So prepossessing step is performed to group (in sequential order) all the frames of same label in single action category. As a result, we are left with 34 different video sequences of 8 different actions, each action having the following number of video sequences (seq.s). The actions are "run left angular (rla)" (5 seq.s), "run left (rl)" (5 seq.s), "walk left (wl)" (3 seq.s), "walk in/out (wio)" (5 seq.s), "run in/out (rio)" (5 seq.s), "walk right (wr)" (5 seq.s), "run right (rr)" (3 seq.s) and "run right angular (rra)" (3 seq.s). Some mistakes are made by the proposed approach because of the ambiguous nature of poses. For example, the algorithm is confused between "rla" and "rl". Some of the poses are quite confusing to distinguish between two actions. Similar explanation holds for "rra" action versus "rr" action.

Number of 'meaningful' poses in Soccer dataset is slightly lower than other datasets because in soccer, pose ambiguity is high and only a handful of unambiguous poses exist in each action class. So the confusion matrix (Table 3) of the proposed approach is obtained by taking fewer number of poses (Table 2) for each of the eight actions in comparison to other datasets. The overall accuracy of the proposed approach for soccer dataset is 82.35%.

3.2 Tower dataset.

The Texas Austin (Tower) dataset for human action recog-

Table 1: Classification accuracy of proposed method compared to the state-of-the-arts

Methods	Overall accuracy (%)			
	Soccer data	Tower data	Hockey data	KTH data
S-LDA	77.81	93.52	87.50	91.20
S-CTM	78.64	94.44	76.04	90.33
Proposed Method using meaningful poses	82.35	97.22	89.58	92.83

Table 2: Classification accuracy of proposed method compared to the accuracy given by some fixed number of poses per action

Datasets	Proposed method with number of selected poses per action							Proposed method using 'meaningful' poses
	1	2	3	4	5	6	7	
Soccer	70.59	73.53	**79.41**	76.47	73.53	67.65	64.71	82.35
Tower	82.41	84.26	90.74	**95.37**	91.67	83.33	77.78	97.22
Hockey	81.25	83.33	**87.50**	83.33	75.00	72.92	70.83	89.58
KTH	86.50	87.67	88.33	90.17	**91.33**	90.33	89.67	92.83

Table 3: Confusion matrix of Soccer dataset (entries given in %)

	rla	rl	wl	wio	rio	wr	rr	rra
rla	80	20	0	0	0	0	0	0
rl	20	80	0	0	0	0	0	0
wl	0	0	100	0	0	0	0	0
wio	0	0	0	80	20	0	0	0
rio	0	0	0	20	80	0	0	0
wr	0	0	0	0	0	100	0	0
rr	0	0	0	0	0	0	67	33
rra	0	0	0	0	0	0	33	67

Table 4: Confusion matrix of Tower dataset (entries given in %)

	P	S	D	W	C	R	W1	W2	J
P	100	0	0	0	0	0	0	0	0
S	0	83	0	0	0	0	0	0	17
D	0	0	92	0	0	8	0	0	0
W	0	0	0	100	0	0	0	0	0
C	0	0	0	0	100	0	0	0	0
R	0	0	0	0	0	100	0	0	0
W1	0	0	0	0	0	0	100	0	0
W2	0	0	0	0	0	0	0	100	0
J	0	0	0	0	0	0	0	0	100

nition consists of 108 video sequences of nine different actions of six different peoples, each people showing each action twice. The nine actions are, "pointing (P)", "standing (S)", "digging (D)", "walking (W)", "carrying (C)", "running (R)", "wave 1 (W1)", "wave 2 (W2)", "jumping (J)". The Tower dataset is actually a collection of aerial action videos where the performer is filmed from tower top and he appears as a tiny blob of height 30 pixels approximately. The approximate bounding rectangles of the human performer as well as foreground filter-masks are supplied with the dataset.

We make use of the bounding rectangle and ignore the foreground filter mask. Since each video clip contains a single action, the video clips are already grouped into respective action classes and we do not need any preprocessing step as we did in case of Soccer data. The rest of the process is essentially same. The confusion matrix (Table 4) illustrates the class wise recognition rate for each action. The proposed approach achieves an overall accuracy of 97.22% on this dataset. We show the confusion matrix for per-video classification of our approach.

3.3 Hockey dataset.

The Hockey dataset consists of 70 video tracks of hockey players with 8 different actions, e.g., "skate down (D)", "skate left (L)", "skate leftdown (Ld)", "skate leftup (Lu)", "skate right (R)", "skate rightdown (Rd)", "skate rightup (Ru)" and "skate up (U)". The confusion matrix is shown in Table 5.

Table 5: Confusion matrix of Hockey dataset (entries given in %)

	D	L	Ld	Lu	R	Rd	Ru	U
D	100	0	0	0	0	0	0	0
L	0	100	0	0	0	0	0	0
Ld	0	0	83	17	0	0	0	0
Lu	0	0	17	83	0	0	0	0
R	0	0	0	0	100	0	0	0
Rd	0	0	0	0	0	67	33	0
Ru	0	0	0	0	0	17	83	0
U	0	0	0	0	0	0	0	100

The proposed approach has acheived an overall accuracy of 89.58% on this dataset. Like soccer, our algorithm finds less number of 'meaningful' poses for each action class in hockey dataset due to increased pose ambiguity. Most of the mistakes done by the proposed approach are reasonable, e.g., our method becomes confused between the actions "Ld" and "Lu", similarly between "Rd" and "Ru".

3.4 KTH dataset.

The KTH dataset of human motion contains six different types of human actions, namely "boxing (B)", "hand clapping (Hc)", "hand waving (Hw)", "jogging (J)", "running

Table 6: Confusion matrix of KTH dataset (entries given in %)

	B	Hc	Hw	J	R	W
B	96	0	0	3	1	0
Hc	0	97	3	0	0	0
Hw	0	0	100	0	0	0
J	1	0	0	86	12	1
R	0	0	0	18	78	4
W	0	0	0	0	0	100

(R)", "walking (W)", performed by 25 different persons for four times each; outdoor, outdoor with scale variation, outdoor with different cloths and indoor. Naturally, most of the confusions occurred for running and jogging because of their almost similar patterns of poses. The overall accuracy of the proposed approach is 92.83%. Table 6 shows the confusion matrix of our method.

4. CONCLUSIONS

This paper studies the action recognition with 'meaningful' poses. From an initial and large vocabulary of poses, the proposed approach prunes out ambiguous poses and builds a small but highly discriminatory codebook of 'meaningful' poses. We demonstrate that identifying 'meaningful' poses can provide vital clue about the kind of human activity. With a sparse descriptor of human poses (and related motion pattern), we build up a histogram of oriented field vectors following a multi-resolution framework. By the notion of centrality theory of graph connectivity we extract the 'meaningful' poses which, we argue, contain semantically important information in describing the action in context. Forming a codebook of 'meaningful' poses, we evaluate our methodology on four standard datasets of varying complexity levels and report improved performance when compared with benchmark algorithms. Presently our algorithm works for single performer; extending it to recognize multiple actions in the same scene may be a future research direction.

5. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] C.-C. Chen, M. S. Ryoo, and J. K. Aggarwal. UT-Tower Dataset: Aerial View Activity Classification Challenge. *http:// cvrc.ece.utexas.edu/ SDHA2010/ Aerial_View_Activity.html*, 2010.

[3] G. K. M. Cheung, S. Baker, C. Simon, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Computer Vision and Pattern Recognition (volume 1)*, pages 77–84. IEEE Computer Society, June 2003.

[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2003.

[5] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, April 2003.

[6] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. Spriger, 2008.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision (volume 2)*, pages 726–733. IEEE Computer Society, October 2003.

[8] L. Fengjun and R. Nevatia. Single view human action recognition using key pose matching and viterbi path seraching. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, 2007.

[9] W. Hoeffding. Probability inequalities for sum of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[10] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Computer Vision and Pattern Recognition*. IEEE Computer Society, July 2008.

[11] W. L. Lu, K. Okuma, and J. J. Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image and Vision Computing*, 27(1/2):189–205, January 2009.

[12] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International Joint Conference on Artificial Intelligence*, pages 674–679. Morgan Kaufmann Publishers Inc., 1981.

[13] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Europian Conference on Computer Vision (volume 3) LNCS 2352*, pages 666–680. Springer, January 2002.

[14] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. In *Computer Vision and Pattern Recognition (volume 2)*, pages 326–333. IEEE Computer Society, June 27-July 2 2004.

[15] B. L. Narayan, C. A. Murthy, and S. K. Pal. Maxdiff kd-trees for data condensation. *Pattern Recognition Letters*, 27(3):187–200, February 2006.

[16] R. Navigli and M. Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692, April 2010.

[17] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, June 2008.

[18] D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning*, pages 727–734. Morgan Kaufmann Publishers Inc., 2000.

[19] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition*, pages 32–36. IEEE Computer Society, 2004.

[20] Y. Wang and G. Mori. Human action recognition by semi-latent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774, October 2009.

[21] D. B. West. *Introduction to Graph Theory*. Prentice Hall, 2000.