Modeling Sense Disambiguation of Human Pose: Recognizing Action at a Distance by Key Poses

Snehasis Mukherjee, Sujoy Kumar Biswas, Dipti Prasad Mukherjee

Electronics and Communication Sciences Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata, India; {snehasismukho, skbhere}@gmail.com, dipti@isical.ac.in

Abstract. We propose a methodology for recognizing actions at a distance by watching the human poses and deriving descriptors that capture the motion patterns of the poses. Human poses often carry a strong visual sense (intended meaning) which describes the related action unambiguously. But identifying the intended meaning of poses is a challenging task because of their variability and such variations in poses lead to visual sense ambiguity. From a large vocabulary of poses (visual words) we prune out ambiguous poses and extract key poses (or key words) using centrality measure of graph connectivity [1]. Under this framework, finding the key poses for a given sense (i.e., action type) amounts to constructing a graph with poses as vertices and then identifying the most "important" vertices in the graph (following centrality theory). The results on four standard activity recognition datasets show the efficacy of our approach when compared to the present state of the art.

1 Introduction

Action recognition at a distance typically happens when the performer is far away from the camera and appears very small (approximately 30 to 40 pixels tall) in the action videos. Modeling the body parts of the performer is difficult since the limbs are not distinctly visible. The only reliable cue in such case is the pose information and we bank on the motion pattern of the poses to derive our descriptors. The contribution in this paper is twofold. First, we propose a novel pose descriptor which not only captures the pose information but also takes the motion pattern of the poses into account. Secondly, our contribution lies in developing a framework for modeling the intended meaning associated with the human poses in different contexts. The activity recognition methodology proposed here is based on the premise that human actions are composed of repetitive motion patterns and a sparse set of key poses (and their related movements) often suffice to characterize an action quite well. The proposed methodology follows the bag-of-word approach and "word" here refers to the pose descriptor of the human figure corresponding to a single video frame. Consequently a "document" corresponds to the entire video of a particular action. The poses can often be very ambiguous and they may exhibit confusing interpretations about the nature of associated actions. The inherent variability present in the poses may make a single pose a likely candidate for multiple action categories or sometimes none at all. Variation in poses is the primary source of visual sense ambiguity and in such cases it becomes difficult to infer which pose signifies what kind of visual senses (i.e., human actions). For example, top row in Fig. 1 shows some ambiguous poses and by looking at them one cannot tell for certain the corresponding actions. Whereas, the bottom row illustrates the key poses which unambiguously specify the related actions. Action recognition in videos by bag-of-word based methods either seek right kind of features for video words [2–5] or model the abstraction behind the video words [6–9]. There are initiatives which study pose specific features [9–11] but modeling visual senses associated with poses in videos is largely an unexplored research area. Our work not only derives novel pose descriptors but, more importantly, seeks 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 [12]. Google uses centrality measures to rank webpages [13] and recently this ranking technique has spelled success in feature ranking for object recognition [14] and video-action recognition [8] tasks. Also, the ambiguity of visual poses bears a direct similarity with the word sense disambiguation [1] in Natural Language Processing (NLP), where the senses associated with a text word vary from context to context and the objective comprises identifying its true meaning expressed in a given context. This paper is all about to get rid of such sense ambiguities associated with different human poses and find out a sparse set of key poses that can efficiently distinguish human activities from each other.

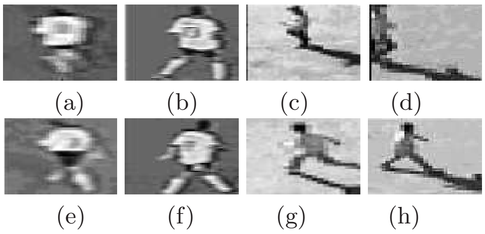


Fig. 1. Top row shows some ambiguous poses (labeled by our algorithm) from Soccer (a-b) and Tower (c-d) datasets and the confusion is in deciding whether they represent running or walking. The bottom row shows retrieved key poses (by our algorithm) for walking (e - f) from Soccer dataset, running (g) and walking (h) from Tower dataset

Section 2 describes the proposed methodology and Section 3 presents results followed by conclusions in Section 4.

2 Proposed Methodology

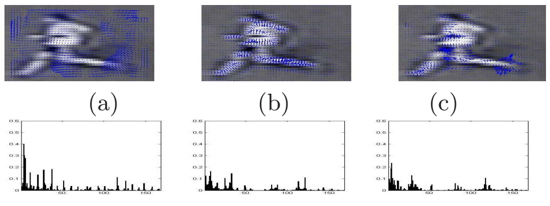
In our approach, each frame of an action video provides a multidimensional vector, which we refer as pose descriptor. The descriptors are derived by combining motion cue from the optical flow field [15] and pose cue from the oriented gradient field of a particular video frame. The pose descriptors, upon clustering (Section 2.2) result into a moderately compact representation S. The key poses are extracted (Section 2.3) from this initial large codebook S in a supervised manner and this sparse set of key poses are used for classification of an unknown target video. We elaborate our descriptor extraction process below.

2.1 Pose Descriptor: Combining Motion & Pose

Our pose descriptors derive the benefit of motion information from optical flow field and pose information from the orientation field. The optical flow field $\overrightarrow{\boldsymbol{F}}$ is weighted with the strength of the oriented gradient field $\overrightarrow{\boldsymbol{B}}$ to produce a resultant flow field that we call $\overrightarrow{\boldsymbol{V}}$, i.e.,

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

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

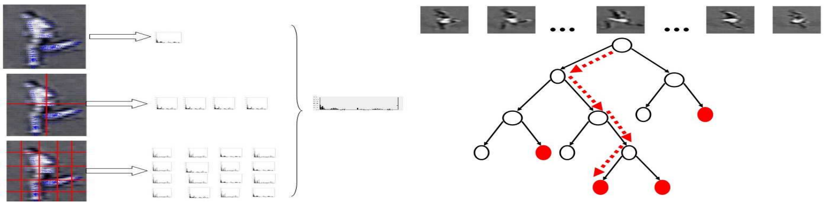

(a) (b) (c)

(d) (e) (f)

Fig. 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))

The effect of this weighted optical flow field is best understood if one treats the oriented gradient field as a band pass filter (Fig. 2). The optical flow vectors, appearing in large magnitude on the background (Fig. 2(a)) quite away from the human figure (originated due to signal noise or unstable motion), get "blurred" and eventually filtered out on modulation with the gradient field. This is because the gradient field takes high value where the edge is prominent, preferably along the edge boundary of the foreground figure, but it is very low in magnitude on the uniform background space (Fig. 2(b)). 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 (Fig. 2(c)). So our descriptor is basically a motion-pose descriptor preserving the motion pattern of the human pose. Following section gives the final details for building up the descriptor.

Suppose we are given a collection of M frames in a video sequence of some action A and we represent the sequences by $I_1, I_2, \ldots, I_M$. We define I to be an $m \times n$ grey image defined as a function such that for any pixel (x, y), where $(x, y) \in Z \times Z$, the image $I(x, y) \in \theta, \theta \subset Z$. Corresponding to I_{t-1} and I_t $(t = 2, 3, \ldots, M)$ we compute the optical flow field $\overrightarrow{\boldsymbol{F}}$. Also, we derive the gradient field $\overrightarrow{\boldsymbol{B}}$ corresponding to frame I_t and following (1) we obtain $\overrightarrow{\boldsymbol{V}}$. We consider a three layer image pyramid (Fig. 3(a)) where in the topmost layer we distribute the field vectors of $\overrightarrow{\boldsymbol{V}}$ 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 next layer in the image pyramid splits the image into 4 equal (or almost 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 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.


(a)

(b)

Fig. 3. (a) Formation of 168-dimensional pose descriptor in three layers. (b) 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 key poses

2.2 Unsupervised Learning of Visual Codebook of Poses

Human activity follows a sequence of pose patterns and such pose sequence occurs in a repetitive fashion throughout the entire action span. The shortest sequence that gets repeated in an entire action video is defined as action cycle C and within an action cycle C let us suppose T poses use to occur where T is the cycle length of the shortest action sequence. For a given action video $A = \{I_1, I_2, \ldots, I_M\}$, we can write a sequence $A = \{C_1, C_2, \ldots, C_H\}$, where $H << M$ and C_i $(i = 1, 2, \ldots, H)$ denotes the action cycle. In action A, each frame I_i $(i = 1, 2, \ldots, M)$ produces an action descriptor q_i (1). Clearly $A = \{C_1, C_2, \ldots, C_H\}$ seeks a true partitioning of the pose descriptors $\{q_i\}_{i=1}^{M}$ for all types of action A. Ideally $C_1 = C_2 = \ldots = C_H$ as all the action cycles should contain same set of pose patterns. But in practice C_i and C_j $(i \neq j; C_i, C_j \in A)$ are not same because of pose variation, related noise, uncertainty related to cycle length T. However, for a given action video A, C_i and C_j together contain redundant information as many of the poses that occur in C_i also occur in C_j. This follows from the cyclical nature of the human action. Clustering technique removes much of such redundancies present in the pose descriptors $\{q_i\}_{i=1}^{M}$. Please note, our intention here is not to seek true partitioning of pose descriptors but to obtain a large vocabulary S where at least some of the redundancies of the pose descriptors are eliminated. An optimum (local) lower bound on the codebook size of S can be estimated by Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) [16] or one can directly employ X-means algorithm [17] which is a divisive clustering technique where splitting decision depends on the local BIC score (i.e., does BIC increase or decrease upon splitting the parent cluster into child ones). K-means based clustering techniques [17] rely on Euclidean distance metric which is isotropic in nature and suffer from curse of dimensionality when the dimension of the feature space increases. To alleviate the curse of dimensionality we adopted a kd-tree based clustering algorithm [18] where the leaf nodes of the kd-tree would indicate our pose clusters. One can also choose (depending on computational expense) multiple samples from each leaf node to construct the large pose vocabulary $S = \{p_1, p_2, \ldots, p_R; p \in \Re^d\}$, where d denotes the dimensionality of the pose descriptors.

2.3 What Does a Pose Tell Us? Modeling Sense Disambiguation of Human Pose

The pose codebook S contains poses which are often ambiguous. We first outline the motivation behind the choice of our model and then explain how we model the visual senses and prune out the ambiguous poses.

Motivation behind Sense Disambiguation of Pose: Fig. 1 shows examples of some poses and the following relations illustrate what sense or action type the poses in Fig. 1 depict. Poses in relation (i) confuse between two senses (i.e., waking and running) but poses in relation (ii) and (iii) strongly express the associated sense. (i) $\{Fig.1(a)\}or\{Fig.1(c)\} \rightarrow \{running, walking\}$; (ii) $\{Fig.1(e)\} \rightarrow \{walking\}$; (iii) $\{Fig.1(g)\} \rightarrow \{running\}$.

We identify the key poses separately for each action type. For each action A we construct a pose graph which captures the very basic essence - how poses in S occur over the entire action span. Each action cycle C in A is defined as a pose sequence, i.e., $C = \{p_1, p_2, \ldots, p_T\}$, $C \in A$ and $p_i \in S$ $(i = 1, 2, \ldots, T)$. When we see a pose p' that is strongly suggestive of some particular action A, we can expect that most of the action cycles C in A contain that pose; for example, Fig. 1(g) illustrates one such key pose which is difficult to miss in any action cycle of a running video. Please note such key pose p' not necessarily has maximum number of occurrences in action video A. Rather such pose must occur at uniform interval, neither only in the beginning nor only at the end and nor intermittently in between. A pose p' which has a strong sense associated (indicating that it represents action A) must have the highest cardinality of the following set $\lambda_{p'|A}$ (set λ of p' given action A) given by

$$\lambda_{p'|A} = \{C | C \in A \text{ } and \text{ } p' \in C\}. \tag{2}$$

Poses which occur all of a sudden or are not prominent in each action cycle can be considered as deviations from the regular poses. Such irregular pose patterns are ambiguous and they need to be pruned out from S. They happen primarily because of tremendous variability of human poses and secondarily, though to a lesser extent, because of associated noise (for example, shadows in tower dataset, transmission noise in soccer dataset).

Construction of Pose Graph G: The pose graph for each action type A contains poses from S as vertices and an edge between two pose vertices explains the joint behavior of two poses - how well they "describe" the action together. By "describe" we of course mean how regularly the pose u (or v) occurs in an action video A. It is quite reasonable that a pose u that occurs regularly in A has a high cardinality $\lambda_{u|A}$. This is because more regularly it occurs, the more action cycles it will belong to. Next we define the pose graph.

Definition 1. *A pose graph for a given action type A is an undirected edge-labeled graph $G = (S, E)$ where each vertex in G corresponds to a pose belonging to the initial 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 $0 < \omega(u, v) \leq 1$. It is assumed that ω is symmetric i.e., $\omega(u, v) = \omega(v, u)$, for all u, v and $\omega(u, u) = 0$ for all u.*

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

where $\eta(u, v) = |\lambda_{u|A} \bigcap \lambda_{v|A}|$.

In practical setting ∞ can be replaced by a large constant when u and v do not have an edge in between them. Construction of G requires computation of $\eta(u, v)$ which again depends on $\lambda_{u|A}$ (2). For computing this set one has to map each video frame (actually the pose descriptor p derived from this frame) to the most appropriate element in S (either by nearest neighbor computation or a suitable Kd-tree traversal algorithm illustrated in Fig. 3(b)).

Key Pose Selection by Pose Ranking: Given the pose graph G for a particular action the task next is to evaluate which one of the poses is most "important" in characterizing an action. An equivalent problem exists in social network analysis [1, 12] (viz. search in matrimonial websites or business/social network websites) where importance of a node (may be a webpage or more specifically a person) in a network is identified using centrality measure of graph connectivity. Such ranking technique is used in web search (Google's PageRank algorithm

[13]) and it has been recently used with success in feature ranking for object recognition [14] and feature mining task [8]. Inspired by their success we seek to rank poses and choose the N-best key poses for a particular kind of action. Grouping all such key poses together we build our discriminatory codebook ξ.

There are various centrality measures of graph connectivity to accomplish the ranking task in a pose graph [1]. Basically the choice of connectivity measure influences the selection process for the highly ranked poses. Given a graph connectivity measure "e" and the set of vertices S and $u, v \in S$, we induce a ranking $rank_e$ of the vertices u and v such that $rank_e(u) \leq rank_e(v)$ iff $e(u) \geq e(v)$. Then for each action type A, we select the best ranking pose v according to $e(v)$. To make e explicit we adopt eccentricity as a measure of graph connectivity [12]. We choose eccentricity because it is simple and fast to compute. Like other centrality measures it relies on the following notion of centrality - a node is central if it is maximally connected to all other nodes. Centrality measure determines the degree of relevance of a single vertex u in a graph G and hence can be viewed as a measure of influence over the network.

Definition 2. *Given a pose graph $G = (S, E)$, where the vertex set is represented by initial codebook S and E stands for edges, 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 a vertex $u \in S$ is the maximum distance from u to any other vertex $v \in S$, i.e., $e(u) = max\{d(u, v)|v \in S\}$.*

Ranking of poses is strictly based on eccentricity score and the implementation is straightforward. The Floyd-Warshall algorithm [19] computes all-pair-shortest path to evaluate the eccentricity $e(u)$ of each pose $u \in S$. For each action, we choose its N-best key poses by selecting poses with N-lowest eccentricity in a pose graph. Once we identify the key poses for a particular kind of action we repeat the same process for all kinds of actions. The key poses $p_1, p_2, \ldots, p_k$ extracted from all the action types are grouped together and serve as key pose codebook ξ

$$\xi = \{p_1, p_2, \ldots, p_k\} \;\; \forall \;\; p_i \in S, \quad where \;\; i = 1, 2, \ldots, k. \tag{4}$$

2.4 Action Descriptor for Target Video Classification

Once we have a video of frames $I_1, I_2, \ldots, I_M$ and the key pose codebook ξ, the traditional bag-of-words implementation requires extracting pose descriptor q_r for each frame I_r and then mapping it to some key pose p_i, $\forall i = 1, 2, \ldots, k$, in a hard way and thereby building the codebook histogram AD (AD stands for action descriptor), where each bin i of AD keeps the occurrence count of key pose p_i, $i = 1, 2, \ldots, k$. The histogram AD will be used for classification task by a suitable classifier. In [20], Gemert *et. al.* showed soft allocation in codebook increases the classification accuracy. We studied the four codebook models (as described in [20]) and documented their performance in the results section.

Fig. 5 shows the effect of the number of key poses on the overall accuracy for different types of codebook model - plausibility, uncertainty, kernel codebook and traditional codebook. Plausibility performs the best among all and we get the best overall accuracy when the average number of key poses selected for each action is, three. Number of key poses in Soccer is slightly lower than other dataset because in soccer, pose ambiguity is high and only a handful of key poses exist in each action class. We use support vector machines (libSVM [24]) for classification of histogram features AD following "one versus all" framework for multi-class classification.

3 Experiments, Results and Discussions

Video Dataset: The choice of dataset is made keeping in mind the focus of our paper - recognizing action at a distance. Soccer [4], tower [21], hockey [22] datasets contain human performer far away from the camera and 40 pixels tall approximately. Only exception is KTH [23] dataset where we evaluated our proposed methodology on medium size (1̃00 pixels tall) human figure. All the action videos are clearly labeled and we used the labels as provided with the dataset. The soccer dataset contains several video sequences of digitized World Cup football game from an NTSC video tape [4]. We arranged this entire dataset into a total of 34 different video sequences of 8 different actions - run left angular, run left, walk left, walk in/out, run in/out, walk right, run right and run right angular. The Texas Austin (tower) dataset for human action recognition consists of 108 video sequences of nine different actions performed by six different people, each person showing every action twice. The nine actions are pointing, standing, digging, walking, carrying, running, wave 1(one hand), wave 2 (both hands), jumping. The bounding rectangles of the human performer (as well as foreground filter-masks which we did not require) are supplied with the dataset. The hockey dataset consists of 70 video tracks of hockey players with 8 different actions, e.g., skate down, skate left, skate leftdown, skate leftup, skate right, skate rightdown, skate rightup and skate up. The KTH dataset of human motion contains six different types of human actions (boxing, handclapping, hand waving, jogging, running, walking) performed by 25 different persons for 4 times each in the following environments - outdoor, outdoor with scale variation, outdoor with different cloths and indoor.

Table 1. Classification accuracy of proposed approach compared to state-of-the-art

Activity	Overall accuracy (%)			
	Soccer data	Tower data	Hockey data	KTH data
S-LDA [6]	77.81	93.52*	87.50	91.20
S-CTM [6]	78.64	94.44*	76.04	90.33
Effros [4]	49.23	–	–	–
Niebles [7]	–	–	–	81.50
Performance over ξ	79.41	95.37	88.50	91.33
Performance over S	58.83	72.22	68.75	72.22

* based on our implementation of [6]

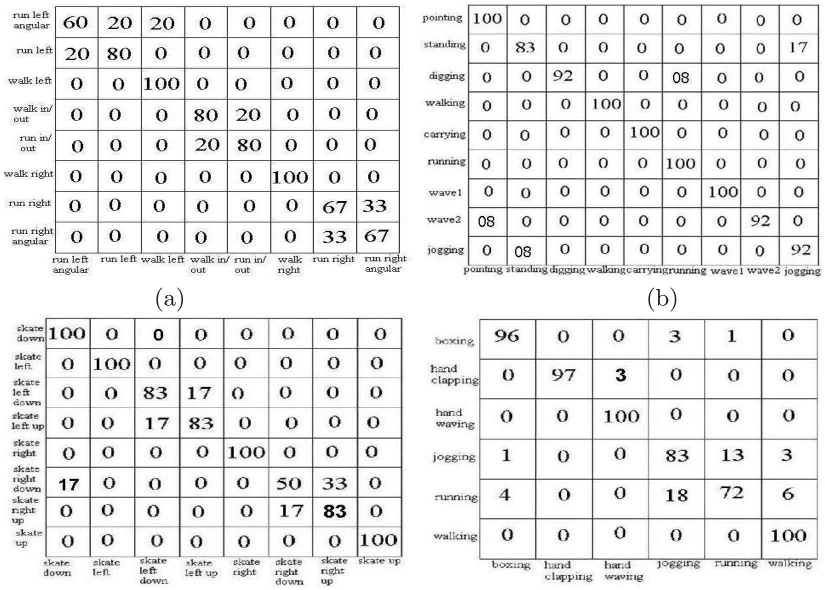
3.1 Experimental Setup

We use support vector machines [24] with radial basis function for classification task following a "Leave-one-out" scheme. A 10-fold cross validation is performed on the training set to tune the parameters of the radial basis function. Holding one sequence out we build our codebooks and action descriptors from the training set and then try to classify the action descriptor of the held out video sequence. This is repeated for all datasets. Please note that used the datasets as available and no preprocessing step is involved. Our approach is efficient both in terms of consumed time and accuracy in detecting human actions. Though building the key pose vocabulary separately for each action takes longest time, but this is done once and we reap benefit later while classifying video with a small set of just 5 or 6 key poses per actions. The average time consumed by key pose dictionary construction amounts to little less than one minute. We assumed the cycle length T constant at 10 because most of the video actions complete a full cycle within 10 frames. The classification task takes a few seconds on a machine with processor speed 2.37 GHz. 512 MB RAM.

3.2 Results & Discussions

We report the overall accuracy of our proposed method and compare them with the state-of-the-art (Table 1). One important achievement of the proposed method is that ξ is built with 5/6 key poses per action. Our codebook is much smaller than the codebook used by [6]. The confusion matrices in Fig. 4 illustrate the class wise recognition rate for each action and it is apparent that our model confuses when two action types have a number of similar poses. In soccer dataset, variation in poses is quite high and it is often difficult (even for humans) to distinguish between two very similar actions (like *run-left* and *run-left-angular*). Consequently some mistakes are made by the proposed approach because of the ambiguous nature of poses. In the Tower dataset, the video quality is relatively better than soccer but shadows are very prominent in the images. We did not remove the shadows; instead we allowed them to provide extra cues about the action type. In hockey video, the major confusion is between left up and left down (or right up and right down). The poses are quite similar in these two actions. In KTH most of the confusions occurred in classifying running and jogging because of their almost similar patterns of poses.

The graphs in Fig. 5 reveal an important fact - with more poses per action recognition accuracy drops down. This is expected as liberal choice of poses from S adds ambiguity to the key pose vocabulary ξ and confuses the recognition task. We build ξ with (on an average) 3, 4, 3, 5 key poses selected from S for soccer, tower, hockey and KTH dataset. Our design decision to use ξ as a refined codebook (built from S) is substantiated by the recognition accuracies with S as our codebook (Table 1). The performance - when S is used in place of ξ for deriving action descriptor AD followed by classification - is much worse suggesting that pruning out ambiguous poses does pay off.



	run left angular	run left	walk left	walk in/ out	run in/ out	walk right	run right	run right angular
run left angular	60	20	20	0	0	0	0	0
run left	20	80	0	0	0	0	0	0
walk left	0	0	100	0	0	0	0	0
walk in/ out	0	0	0	80	20	0	0	0
run in/ out	0	0	0	20	80	0	0	0
walk right	0	0	0	0	0	100	0	0
run right	0	0	0	0	0	0	67	33
run right angular	0	0	0	0	0	0	33	67

(a)

	pointing	standing	digging	walking	carrying	running	wave1	wave2	jogging
pointing	100	0	0	0	0	0	0	0	0
standing	0	83	0	0	0	0	0	0	17
digging	0	0	92	0	0	08	0	0	0
walking	0	0	0	100	0	0	0	0	0
carrying	0	0	0	0	100	0	0	0	0
running	0	0	0	0	0	100	0	0	0
wave1	0	0	0	0	0	0	100	0	0
wave2	08	0	0	0	0	0	0	92	0
jogging	0	08	0	0	0	0	0	0	92

(b)

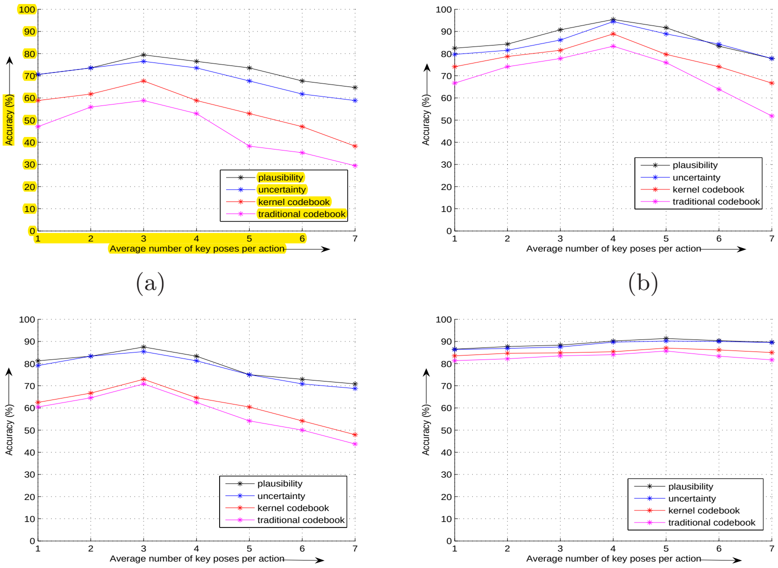
	skate down	skate left	skate left down	skate left up	skate right	skate right down	skate right up	skate up
skate down	100	0	0	0	0	0	0	0
skate left	0	100	0	0	0	0	0	0
skate left down	0	0	83	17	0	0	0	0
skate left up	0	0	17	83	0	0	0	0
skate right	0	0	0	0	100	0	0	0
skate right down	17	0	0	0	0	50	33	0
skate right up	0	0	0	0	0	17	**83**	0
skate up	0	0	0	0	0	0	0	100

(c)

	boxing	hand clapping	hand waving	jogging	running	walking
boxing	96	0	0	3	1	0
hand clapping	0	97	**3**	0	0	0
hand waving	0	0	100	0	0	0
jogging	1	0	0	83	13	3
running	4	0	0	18	72	6
walking	0	0	0	0	0	100

(d)

Fig. 4. Accuracy plot with average number of key poses per action for (a) Soccer dataset, (b) Tower dataset, (c) Hockey dataset and (d) KTH dataset



(a)

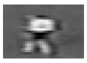
(b)

(c)

(d)

Fig. 5. Accuracy plot with average number of key poses per action for (a) Soccer dataset, (b) Tower dataset, (c) Hockey dataset and (d) KTH dataset

Table 2. Key poses of Soccer dataset (4 actions) with corresponding eccentricity value

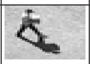
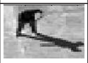
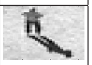
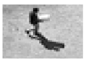
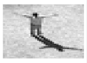
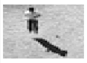
Activity	rla	rl	wl	rio
Key poses (top three key poses from each activity) with corresponding eccentricity value				
	1.89	2.0	1.73	1.42
	2.10	2.5	1.98	1.49
				
	2.33	2.5	2.15	1.67

Tables 2 and 3 show some key poses retrieved by our algorithm along with their centrality measures. Each of the key poses clearly indicates the intended sense and the related human action becomes immediately apparent.

4 Conclusions and Future Scope

This paper studies the problem of key pose retrieval for definite action patterns by modeling the visual senses expressed by different human poses. From a large vocabulary of poses the proposed methodology prunes out ambiguous poses and builds a small but highly discriminatory codebook of key poses. In selecting the key poses we made a ranking of poses for a given action type using centrality theory and choose N-best poses among them. The reported accuracy with our small codebook size is slightly superior to the state of the art. It is demonstrated that identifying key poses can provide vital clue about the kind of human activity.

Table 3. Key poses of Tower dataset (4 actions) with corresponding eccentricity value

Activity	C	D	W2	J
Key poses (top three key poses from each activity) with corresponding eccentricity value				
	1.27	1.47	1.5	1.3
	1.29	1.59	1.52	1.45
				
	1.29	1.72	1.6	1.49

References

1. Navigli R., Lapata M.: An experimental study of graph connectivity for unsupervised word sense disambiguation. IEEE Trans. on PAMI **32(4)** (2010) 678–692
2. Dollar P., Rabaud V., Cotrell G., Belongie S.: Behavior Recognition via Sparse Spatio-Temporal Features. *IEEE Int. Workshop on VS-PETS* (2005) 65–72
3. Laptev I., Lindeberg T.: Space-time Interest Points. *9th ICCV* **1** (2003) 432–439
4. Efros A.A., Berg A. C., Mori G., Malik J.: Recognizing Action at a Distance. *9th ICCV.* **2**, (2003) 726–733
5. Ikizler N., Duygulu P.: Histogram of Oriented Rectangles: A New Pose Descriptor for Human Action Recognition. Image and Vision Computing **27** (2009) 1515–1526
6. Wang Y., Mori G.: Human Action Recognition by Semi-Latent Topic Models. IEEE Trans. on PAMI **31(10)** (2009) 1762–1774
7. Niebles J.C., Wang H., Li Fei-Fei: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. IJCV **79(3)**, (2008) 299–318
8. Liu J., Luo J., Shah M.: Recognizing Realistic Actions from Videos “in the Wild”. *CVPR* (2009)
9. Niebles J., Le Fei Fei: A hierarchical model of shape and appearance for human action classification. *CVPR* (2007)
10. Bissacco A., Yang M. H., Soatto S.: Detecting humans with their pose. *NIPS* (2007)
11. Fengjun L., Nevatia R.: Single View Human Action Recognition using Key Pose Matching and Viterbi Path Seraching. *CVPR* (2007)
12. Wasserman S., Faust K.: Social Network Analysis: Methods and Applications. *Cambridge University Press* (1994)
13. Brin S., Page L.: The anatomy of a large-scale hyper-textual web search engine. *Computer Networks and ISDN Systems* **30(1-7)** (1998) 107–117
14. Kim G., Faloutsos C., Hebert M.: Unsupervised modeling of object categories using link analysis technique. *CVPR* (2008)
15. Lucas B.D., Kanade T.: An Iterative Image Registration Technique with an Application to Stereo Vision. *7th IJCAI* (1981) 674–679
16. Bishop C. M.: Pattern Recognition and Machine Learning. *Springer* (2006)
17. Pelleg D., Moore A. W.: X-means: Extending K-means with efficient Estimation of the Number of Clusters. *ICML* (2000)
18. Narayan B.L., Murthy C.A., Pal S.K.: Maxdiff kd-trees for Data Condensation. PRL **27(3)** (2005) 187–200
19. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C.: Introduction to Algorithms. *MIT Press* (2003)
20. Gemert J.C.V., Veenman C.J., Smeulders A.W.M., Geusebroek J.M.: Visual Word Ambiguity. IEEE Trans. on PAMI **32(7)** (2010) 1271–1283
21. Chen C.C., Ryoo M.S., Aggarwal J.K.: UT-Tower Dataset: Aerial View Activity Classification Challenge, Year 2010, http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html
22. Lu W.L., Okuma K., Little J.J.: Tracking and Recognizing Actions of Multiple Hockey Players Using the Boosted Particle Filter. Image and Vision Computing **27(1-2)** (2009) 189–205
23. Schuldt C., Laptev I., Caputo B.: Recognizing Human Actions: A Local SVM Approach. *17th ICPR* (2004) 32–36
24. http://www.csie.ntu.edu.tw/ cjlin/libsvm/ (June 2010)