Recognizing Conversational Interaction Based on 3D Human Pose

Jingjing Deng, Xianghua Xie*, Ben Daubney, Hui Fang, and Phil W. Grant

Department of Computer Science, Swansea University,
Singleton Park, Swansea SA2 8PP, United Kingdom
x.xie@swansea.ac.uk
http://csvision.swan.ac.uk

Abstract. In this paper, we take a bag of visual words approach to investigate whether it is possible to distinguish conversational scenarios from observing human motion alone, in particular gestures in 3D. The conversational interactions concerned in this work have rather subtle differences among them. Unlike typical action or event recognition, each interaction in our case contain many instances of primitive motions and actions, many of which are shared among different conversation scenarios. Hence, extracting and learning temporal dynamics are essential. We adopt Kinect sensors to extract low level temporal features. These features are then generalized to form a visual vocabulary that can be further generalized to a set of topics from temporal distributions of visual vocabulary. A subject-specific supervised learning approach based on both generative and discriminative classifiers is employed to classify the testing sequences to seven different conversational scenarios. We believe this is among one of the first work that is devoted to conversational interaction classification using 3D pose features and to show this task is indeed possible.

Keywords: 3D human pose, conversational interaction classification, interaction analysis, Kinect sensor.

1 Introduction

Human action and activity recognition has proved to be viable in video surveillance applications throughout the years [11,1,18], though it still remains an open and challenging problem. There is however already a body of work interested in the detection and recognition of social interaction between multiple people [5,7], which is particularly difficult since the actions of multiple subjects must be inferred and understood.

From the feature selection perspective, both low-level appearance features, such as color, dense optical flow, spatio-temporal interest point, and high-level human pose features have been investigated. However, initially, the dependence on low-level features has meant that the class of social interactions examined

* Corresponding author.

J. Blanc-Talon et al. (Eds.): ACIVS 2013, LNCS 8192, pp. 138–149, 2013.

Fig. 1. Examples of observations made for each pair during different conversational interactions. The time difference between each consecutive frame shown is two seconds. For better visualization, only upper body is shown.

thus far typically have been limited to those that can be readily identified and most easily described by a particular set of motions or poses, e.g. handshake or high-five. Alternatively, observation is made at a coarse level to recognize interactions, which are only dependent on high-level tracking of entire individuals, e.g. in a surveillance setting. Furthermore, Yao *et al.* [2] have shown that pose-based features outperform low-level appearance features to some extent in the short-time action recognition task. However, the estimation of human pose, particularly in 3D that is considered as a strong cue to action and activity recognition, is problematic and inaccurate, which directly leads to little attention to the pose-based action and activity recognition methods in last decades.

In this work, we propose to leverage recent advances in technology in extracting 3D pose using a consumer sensor (Microsoft Kinect) to examine the feasibility of detecting much more high-level behavioral interactions between two people. Rather than recognizing just key social events, we attempt to analyze and detect different conversational interactions. We investigate whether just by observing the 3D pose of two interacting people we can recognize the type of conversation they are conducting. This work is in part motivated by recent work that showed features derived from 3D human pose are much more discriminative than their low-level image based counterparts e.g. [2]. Therefore, we believe that having access to these features provides the capacity of detecting and classifying much more subtle interactions than currently possible. Often the differences between the interactions examined in this work are not themselves intuitive. Moreover, there are large variations among individuals when performing the same task. Hence, our emphasis in this work is to classify, in a *subject-specific* supervised fashion, short clips of conversational interactions into seven different categories that are defined based on individual tasks, such as debate a topic and problem solving, rather than primitive interactions, such as monologue and exchange. Each clip in our case may contain multiple primitive interaction types. We examine the extent of the visual cues provided by humans in recognizing conversational interactions. We thus employ discriminative methods to carry out the classification. In addition, we apply a generative method based on Hidden Markov Model (HMM), which is a popular choice for recognizing sequential action and activity through modeling the dynamics with varying temporal duration, e.g. [15,14,9,16,17]. A coupled version of HMM is also used to explicitly model the interactions. We recognize that generalizing conversational scenarios across subjects is far more challenging than discriminating them. However, this work is useful in understanding the role of bodily movement in conversational interaction and is a necessary step towards generic, non-subject specific modeling. We believe this to be the first work devoted to conversational interactions where we are interested in identifying the content of a conversation using pose features.

2 Data Set Acquisition

Data was collected using a two-Kinect set-up, each person was recorded using a Kinect Sensor, which captured pose at 30fps. Each of the cameras was slightly

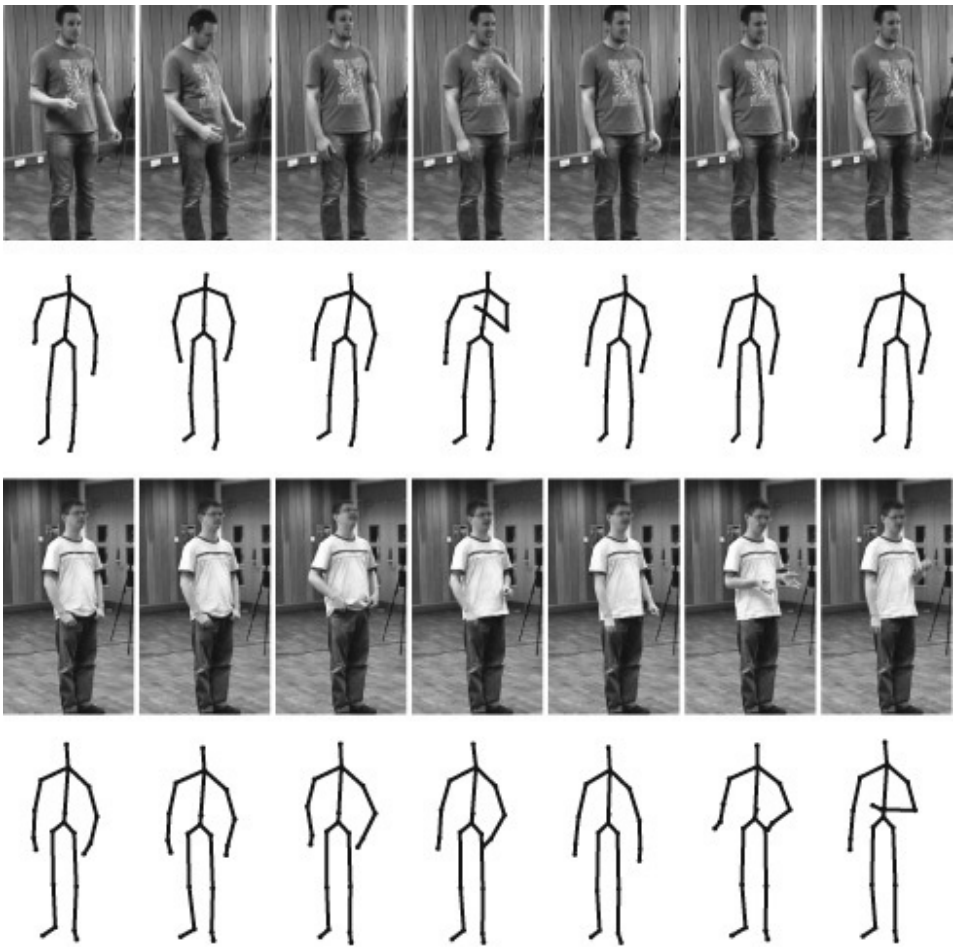

Fig. 2. Example 3D poses from a pair during "Describing Work". Note that the RGB images were captured by separately synchronized cameras at different viewing angles to Kinect - hence the discrepancy in pose. The RGB data is not used in this study.

offset from a direct frontal view so that the participants did not occlude one another. The participants were given seven tasks to complete. The first task was to discuss an area of their current work. The second task was to prepare an interesting story to tell their partner, such as a holiday experience. The third task was to jointly find the answer to a problem. The fourth task was a debate, where the participants were asked to prepare arguments for a particular point of view on an issue we gave to them. In the fifth task they were asked to discuss between them the issues surrounding a statement and come to agreement whether they believe the statement is true or not. The sixth task was to answer a subjective question, and the seventh task was to take it in turn telling jokes to one another. A full description of the different tasks are provided in Table 1.

Each set of seven tasks took about 50 minutes. They were told roughly how long each task to take as a guide, however, they were not being timed or interrupted. Before each task, there were given the opportunity to reread any

Table 1. Description of each of the tasks given to the participants to perform

#	Task Name	Description
1	Describing Work	Each participant was asked to describe to their partner their current work or a project they have involved with. Following this each participant then repeated it back so as to confirm they had understood.
2	Story Telling	Each participant was asked to think of an interesting story they could tell their partner, such as a holiday experience or an experience of a friend.
3	Problem Solving	The participants were given a problem they were asked to think of the solution of together. The problem was "Do candles burn in space and if so what shape and direction?".
4	Debate	The participants were asked to prepare arguments for a given point of view on the topic "Should University education be free?" and then debate this between them.
5	Discussion	The participants were asked to jointly discuss the issues surrounding a statement and come to agreement whether they believe the statement is true or not. The statement was "Social Networks have made the world a better place?"
6	Subjective Question	The participants were asked to discuss a subjective question which was "If you could be any animal, what animal and why?"
7	Telling jokes	The participants were asked to take it turn telling jokes to one another, each participant was provided with three different jokes to learn before attending.

associated material with the task that they may have forgotten. At the end of the session, participants were generally surprised by how much time had passed. A sample of the data collected for each conversational interaction is presented in Fig 2, and the whole dataset is available for download from the following link[1].

3 Proposed Method

As there is no well-defined primitive action or activity categories for conversational interaction, and the gestures vary with different subjects, it is unrealistic to manually annotate the data set. Inspired by the works [13,8,2], the unlabeled low-level features are generalized as a bag of visual words, based on which high-level conversational interaction classification is carried out. The low-level 3D pose features are extracted directly from kinematic human model. Gaussian Mixture Models (GMMs) are fitted to the low-level feature space, and the Gaussian components constitute the vocabulary of visual words. A further generalization of visual words to higher level topics is also investigated. Both discriminative and generative models are trained and applied to recognize the class of unknown sequences for each pair of subjects. The flowchart shown in Fig. 3 illustrates the

[1] `http://csvision.swan.ac.uk/converse.html`

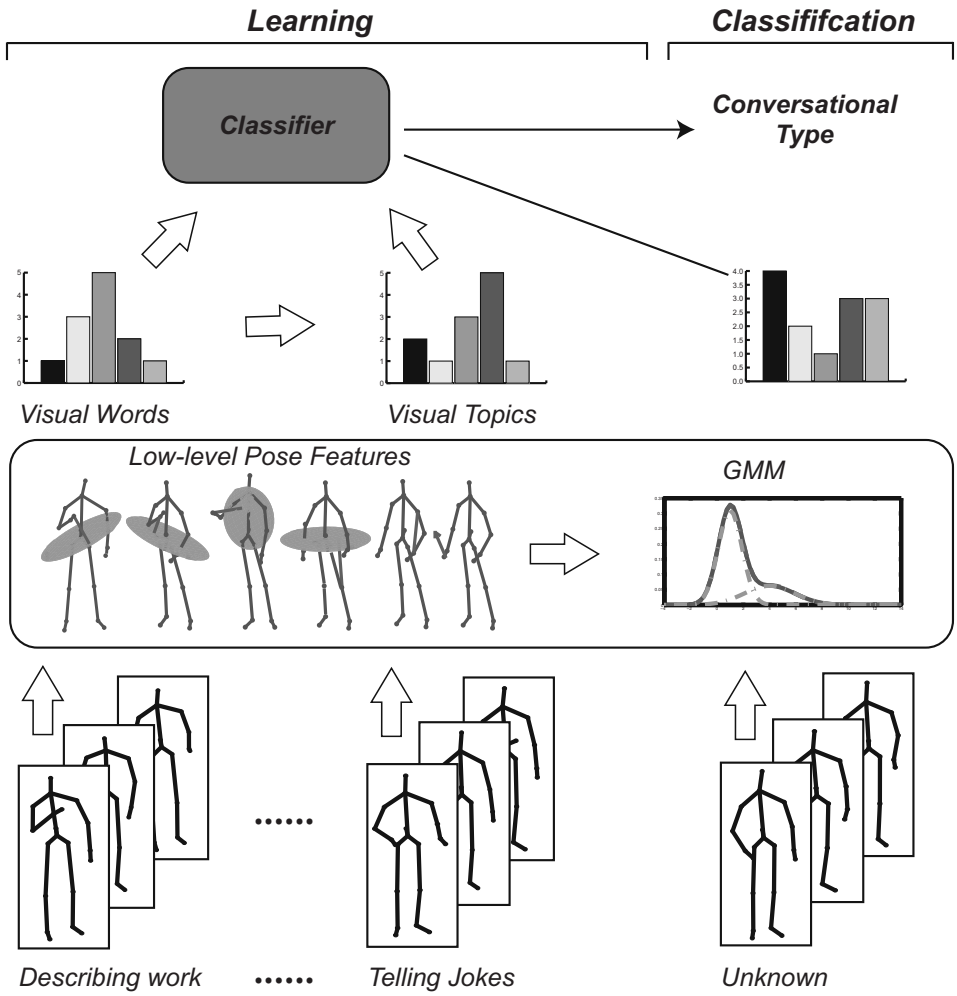

Fig. 3. Flowchart of the proposed method

steps from pose feature extraction, to unsupervised clustering and generalization, and to supervised classification.

3.1 Low-Level Pose Feature

3D pose features have been shown to be useful in motion capture data retrieval and action recognition. Motivated by existed work, such as [2,10,12], we extract three types of features to depict the pose and motion of the upper body. These geometry features extracted from a kinematic chain are simple but powerful for representing human gesture and motion over time. The first set of feature measure the distance between a joint and a reference plane defined using different parts of the body (see Fig. 4(b,c,d,e)). The second set of feature we use measure the distance between two joints at different time intervals and is depicted in Fig. 4(d). The third set of feature measure the velocity of individual joints (see Fig. 4(g)).

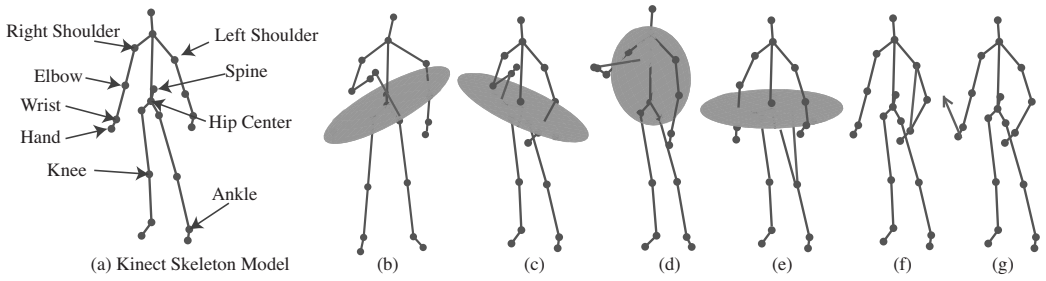

Fig. 4. Visualization of the pose-based features. (a) illustrates the Kinect skeleton model. (b), (c) (d), & (e) shows four reference planes. (f) illustrates the distance between two joints. (g) illustrates a velocity feature at the right hand joint.

There are four reference planes used to quantify the movement of certain joints in the kinematic chain. The first two reference planes are used to measure the distance and velocity of joints on the lower arms, i.e. hands, wrists and elbows. Both planes are located at the same spine point. One of the two planes is defined by the vector connecting the spine and left shoulder (Fig. 4(b)), and the other is defined by the vector connecting the spine and right shoulder (Fig. 4(c)). The former is used to measure the lower arm joints on the left side and the latter is for right side. The two vectors connecting hip center from two shoulders define the third reference plan (Fig. 4(d)), which is used to measure movements of lower arm joints from both arms. The fourth plan is perpendicular to the third plan and crossing the same spine point (Fig. 4(e)). This reference plan is used to measure movement of knees and ankles (ankle points are more stable than feet in Kinect estimation). The overlapping in measurement is to make sure that the 3D motion of those joints are captured among those 2D measurement combinations. Next, we provide the definition for each measurement of joint movement.

The 3D location of a joint at time slice t is denoted as $\omega_{i,t} \in R^3$ and the vector defined by two joints by $\pi_{ij,t} \in R^3$, where i and j indicates the identity of the joints. We define two types of plane $\phi_{ijk,t}$ which is spanned by the joints $\omega_{i,t}, \omega_{j,t}, \omega_{k,t}$, and the plane $\psi_{ijk,t}$ passing through $\omega_{k,t}$ and whose normal vector is aligned with $\pi_{ij,t}$. The normal vector of the plane $\phi_{ijk,t}$ can also be represented by $\pi_{ijk,t}$.

The feature F_d representing the Euclidean distance between joints over Δt is defined as: $F_d = D\{(\omega_{i,t}), (\omega_{j,t+\Delta t})\}$. If $i = j$, then the feature measures the distance of movement of the joint over time Δt, otherwise, it measures the distance between two different joints separated by time.

The features F_{pd1} and F_{pd2} measure the shortest distance from joint $\omega_{n,t}$ to the plane $\phi_{ijk,t+\Delta t}$ and the plane $\psi_{ijk,t+\Delta t}$, respectively, which are defined as: $F_{pd1} = D\{(\omega_{n,t}), (\phi_{ijk,t+\Delta t})\}$ and $F_{pd2} = D\{(\omega_{n,t}), (\psi_{ijk,t+\Delta t})\}$

We also extract F_{jv}, F_{pv}, the component of the joints' velocity along the direction of the vector $\pi_{ij,t+\Delta t}$ and vector $\pi_{ijk,t+\Delta t}$, respectively, which are defined as: $F_{jv} = V\{(\omega_{n,t}), (\pi_{ij,t+\Delta t})\}$ and $F_{pv} = V\{(\omega_{n,t}), (\pi_{ijk,t+\Delta t})\}$. Furthermore, we estimate head orientation from RGB camera output based on the face

localization method [4] and parametric head pose estimation technique [6]. Thus, 49 different low-level pose features are extracted from the Kinect data, with $\Delta t = 1.0s$.

3.2 Middle-Level Visual Words and Topics

The extracted low-level pose features are direct measurements of relative motion at a short time window. Although similar features have been found powerful in classifying primitive actions with short time span [2], what kind of feature is appropriate choice for conversational scenario classification is still an undetermined question, as we cannot decide the conversational scenarios two people conducting just based on the short-term motions. In this work, we adopt the bag of words approach to derive mid-level features that are suitable for classification of conversational interactions, each of which may contain various amount of primitive actions. Different from video analysis where for instance the spatial-temporal interesting points are detected from sequential images using space-time corner detectors or separable linear filters, in our case, the raw data is the locations of joints in the kinematic model. Consequentially, we are concerned with the distributions of those geometrical features across time. We hence using unsupervised clustering to generate visual words across the whole sequence and across all subjects to create a visual vocabulary. A further generalization to visual topics is then performed based on the distribution of visual words in an extended time span that is often larger than typical primitive action.

In information retrieval and natural language processing, the Latent Dirichlet Allocation (LDA) model has been widely used to discover abstract "topics" from a collection of words or low level features. Niebles *et. al.* [13] applied the LDA model to extract action categories from low-level spatial-temporal words in an unsupervised fashion. Inspired this work, we extract the visual words and topics from the low-level pose features. Firstly, a visual vocabulary was constructed by fitting GMMs to each dimension of the low-level pose feature space. We consider each Gaussian component as a visual word. Then, we further assume that those visual words are generated by a mixture of visual topics. To learn those visual topics, we split the sequences into 600 frames sections each of which is considered as a visual document that contains multiple visual topics. The LDA model [3] with a fixed number of latent topics is then applied to all documents, and assigns each visual word in the documents to a potential topic. Next, we use the distributions of those visual words and topics to classify different conversational scenarios.

3.3 Classifiers

Both discriminative classifier and generative classifier are employed in this work, namely Support Vector Machine (SVM), Random Forest (RF) and HMMs. SVM and RF are popular discriminative models for supervised classification. The main reason of choosing them is that they are effective tool to evaluate the discriminative power of our features. Meanwhile, the generative model, HMM could

provide us another perspective in understanding the process of conversational interactions, as it is suitable for modeling the dynamics in the sequential data.

Based on two different middle-level descriptors, visual words and topics, two set of classifiers were trained independently. To train and test the classifiers, each recorded sequence is split into 500 frames sections. Each section is labeled as the task from which it is extracted and used as a single example, both for training and testing. For the discriminative models, the histogram of visual words or topics is computed, and used as feature vector for each section. We learn a random forest with 100 decision trees by randomly sampling with replacement from the complete training set. An SVM with $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ as its kernel function is trained on the same training set. For the HMM, the feature vector of each frame in the section corresponds to an observation node expanded across time. We learn separate HMMs for each of the seven conversational classes. Whilst HMMs are well suited to classifying sequences of different lengths, training a HMM on the section with 150 observation nodes is computationally expensive. Whereas we do not want to inadvertently introduce any bias into the results as because of differing temporal lengths. We down-sample each section by the factor of 12, so that each 150 frames sequence was composed of observations from 13 time instances. As the same as the way we train the discriminative model, two sets of HMM models are learned based on two different middle-level descriptors, i.e. visual words and visual topics. However, a better approach to encode interaction between two subjects using HMM is to use separate HMMs to represent each person and then adding an edge between the two persons across time to build a Coupled HMM (CHMM) [15]. Hence, compared to feature concatenation, CHMM more explicitly model interactions between two subjects.

4 Experimental Result

The approach described in Section 2 was used to collect the data set used in the presented experiments. In total all tasks were completed by 5 different pairs of people, which resulted in more than 500,000 frames. Each class is not obviously distinct from the others, and although there are some representative poses of each class it would be extremely difficult to determine the class using only pose from a single frame. Another major challenge of the data set is the sheer variation in the types of motion and gestures performed by each participant during the task. Even the neutral pose of each participant as they are listening is very different. These make it very difficult for generative methods to classify. The whole dataset is available for download from the following link[2].

To carry out the classification, 10-fold *subject-specific* cross validation is adopted, that is all the sequences were sequentially chopped into 10 segments so that neighboring samples are not distributed across training set and testing set. All the classifier were trained on the same training set independently.

[2] `http://csvision.swan.ac.uk/converse.html`

Table 2. Average subject-specific classification results using the features from only one participant

	Original features	Visual words			Visual topics		
	HMM	HMM	SVM	RF	HMM	SVM	RF
Describing Work	0.79	0.79	0.73	0.84	0.83	0.77	0.80
Story Telling	0.61	0.55	0.49	0.49	0.39	0.65	0.66
Problem Solving	0.50	0.24	0.34	0.27	0.11	0.79	0.79
Debate	0.55	0.33	0.43	0.82	0.12	0.63	0.60
Discussion	0.60	0.48	0.49	0.52	0.36	0.57	0.58
Subjective Question	0.34	0.11	0.25	0.08	0.03	0.67	0.71
Jokes	0.52	0.25	0.34	0.19	0.04	0.57	0.62
Average	0.56	0.39	0.44	0.39	0.27	0.66	0.68

We first test the pose features from only a single person, that is to understand how much information can be extracted by observing one participant in order to determine the topic of their conversation. Table 2 shows the average performance for each method in classifying the seven scenarios using visual words and visual topics as the discriminative feature. When using visual words, an average of 44% and 39% were achieved by SVM and RF classifiers, respectively. When using visual topics, which produces significantly shorter feature vectors (25 vs 340), their performances were increased to 66% and 68%, respectively (see Table 2). This was a significant performance increase for SVM and RF classifiers when using visual topics. It is however notable that HMM performed poor with both features, 36% and 27% compared to a random chance of 14% and 56% accuracy when using original pose features. A reasonable explanation to this is that there are lots of similarities among different scenarios which leads to similarities in low-level features and mid-level descriptors. While generalizing the individual conversation scenario, at current setting HMM emphasized the commonality in the data and hence compromised its discriminative power. This implies that detecting rare events and actions may help the generalization as they are likely more discriminative. It also shows the subtlety in the data set and perhaps large individual variation as well.

For the next experiment we combine features from two participants by concatenating their features before feeding into the classifiers. The results are shown in Table 3. There were broad improvements reported by all three classifiers. For the HMM method, we built a coupled model so that each subject corresponds to a single HMM and both were linked together. The visual words performance improved from 39% to 43%, and from mere 27% to 34% for visual topics. The SVM and RF reported somewhat greater increase in performance, with best result of 76% achieved by RF using visual topics. This clearly highlights the benefit of having multiple streams of information when observing people during an interaction as they can be used to better discriminate the task being performed.

The results we have achieved suggested that it is possible to classify conversational interactions just based on human poses alone for individual pair of

Table 3. Average subject-specific classification results using the features pair two participant

	Original features	Visual words			Visual topics		
	CHMM	CHMM	SVM	RF	CHMM	SVM	RF
Describing Work	0.85	0.85	0.78	0.90	0.92	0.86	0.87
Story Telling	0.78	0.60	0.66	0.52	0.49	0.71	0.77
Problem Solving	0.56	0.19	0.52	0.32	0.16	0.81	0.86
Debate	0.56	0.40	0.59	0.46	0.19	0.65	0.67
Discussion	0.77	0.64	0.63	0.63	0.54	0.63	0.69
Subjective Question	0.31	0.15	0.38	0.17	0	0.75	0.80
Jokes	0.44	0.19	0.51	0.23	0.06	0.64	0.67
Average	0.61	0.43	0.58	0.46	0.34	0.72	0.76

subjects. The generalization of pose interactions, as expected, is a harder problem than discriminating among each others. Whilst the Kinect sensor permits direct estimation of 3D pose that is currently more robust and accurate than RGB camera methods, the data collected still contains some noise, as does the features extracted. However, despite this we have shown that recognition of conversational interactions with subtle differences can still be achieved with high accuracy. More participant data is necessary to analyze the effectiveness of generalized features, and this is leading to a new type of interaction analysis.

5 Conclusion

We presented a comprehensive study on gesture cues in understanding human conversational activity. The difference among the seven scenarios are rather subtle, and the primitive actions and interactions are commonly exhibited across different scenarios. Middle level motion descriptor were generalized from low level pose features obtained from Kinect output. Both discriminative model and generative model were investigated in order to classify subject-specific different types of conversational interactions. It is evident that good classification accuracy can be achieve using discriminative methods. The results also suggests that it is possible to distinguish conversational topic based on the pose movement from a single person. It is however more challenging to generalize different scenarios across subjects. An even larger data set and perhaps more sophisticated HMM models would improve the performance. However, we believe this work offer a somewhat different perspective to action and interaction analysis.

References

1. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. 43(3), 16 (2011)
2. Yao, A., Gall, J., Fanelli, G., Gool, L.V.: Does human action recognition benefit from pose estimation? In: Proceedings of the British Machine Vision Conference, pp. 67.1–67.11. BMVA Press (2011)

3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
4. Fang, H., Deng, J., Xie, X., Grant, P.W.: From clamped local shape models to global shape model. In: Proceedings of the 2013 International Conference on Image Processing, ICIP (2013)
5. Fathi, A.: Social interactions: A first-person perspective. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR 2012, pp. 1226–1233. IEEE Computer Society, Washington, DC (2012), http://dl.acm.org/citation.cfm?id=2354409.2354936
6. Gee, A.H., Cipolla, R.: Determining the gaze of faces in images. Image and Vision Computing 12, 639–647 (1994)
7. Holte, M.B., Tran, C., Trivedi, M.M., Moeslund, T.B.: Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. IEEE Journal of Selected Topics in Signal Processing 6(5), 538–552 (2012)
8. Hospedales, T., Gong, S., Xiang, T.: Video behaviour mining using a dynamic topic model. International Journal of Computer Vision, 1–21 (2012)
9. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 852–872 (2000)
10. Kovar, L., Gleicher, M.: Automated extraction and parameterization of motions in large data sets. ACM Trans. Graph. 23(3), 559–568 (2004)
11. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding 104(2-3), 90–126 (2006)
12. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. ACM Trans. Graph. 24(3), 677–685 (2005)
13. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79(3), 299–318 (2008)
14. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Comput. Vis. Image Underst. 96(2), 163–180 (2004), http://dx.doi.org/10.1016/j.cviu.2004.02.004
15. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 831–843 (2000)
16. Ryoo, M.S., Aggarwal, J.K.: Semantic representation and recognition of continued and recursive human activities. Int. J. Comput. Vision 82(1), 1–24 (2009)
17. Ryoo, M.S., Aggarwal, J.K.: Stochastic representation and recognition of high-level group activities. Int. J. Comput. Vision 93(2), 183–200 (2011)
18. Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Techn. 18(11), 1473–1488 (2008)