A Bag of Words Approach to 3D Human Pose Interaction Classification with Random Decision Forests

Jingjing Deng, Xianghua Xie*, Ben Daubney
Department of Computer Science, Swansea University, Swansea, UK
*x.xie@swansea.ac.uk
http://csvision.swan.ac.uk

Abstract

In this work, we investigate whether it is possible to distinguish conversational interactions from observing human motion alone, in particular gestures in 3D. We adopt Kinect sensors to obtain 3D displacement and velocity measurements, followed by wavelet decomposition 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 supervised learning approach based on Random Forests is used to classify the testing sequences to seven different conversational scenarios. These conversational scenarios 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. That is the interactions we are concerned with are not micro or instant events, such as hugging and high-five, but rather interactions over a period of time that consists rather similar individual motions, micro actions and interactions. 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.

1. Introduction

Human motion capture and activity recognition have proved viable in, for example, computer graphics, media production, robotics, and video surveillance applications throughout the years [21, 16, 1, 23, 5, 22, 14], 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 [7, 10], which is particularly difficult since the actions of multiple subjects must be inferred and understood. However, advances in interaction modeling is of great interest to computer graphics and visual media production.

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 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. Hence, our emphasis in this work is to classify, in a 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 solv-

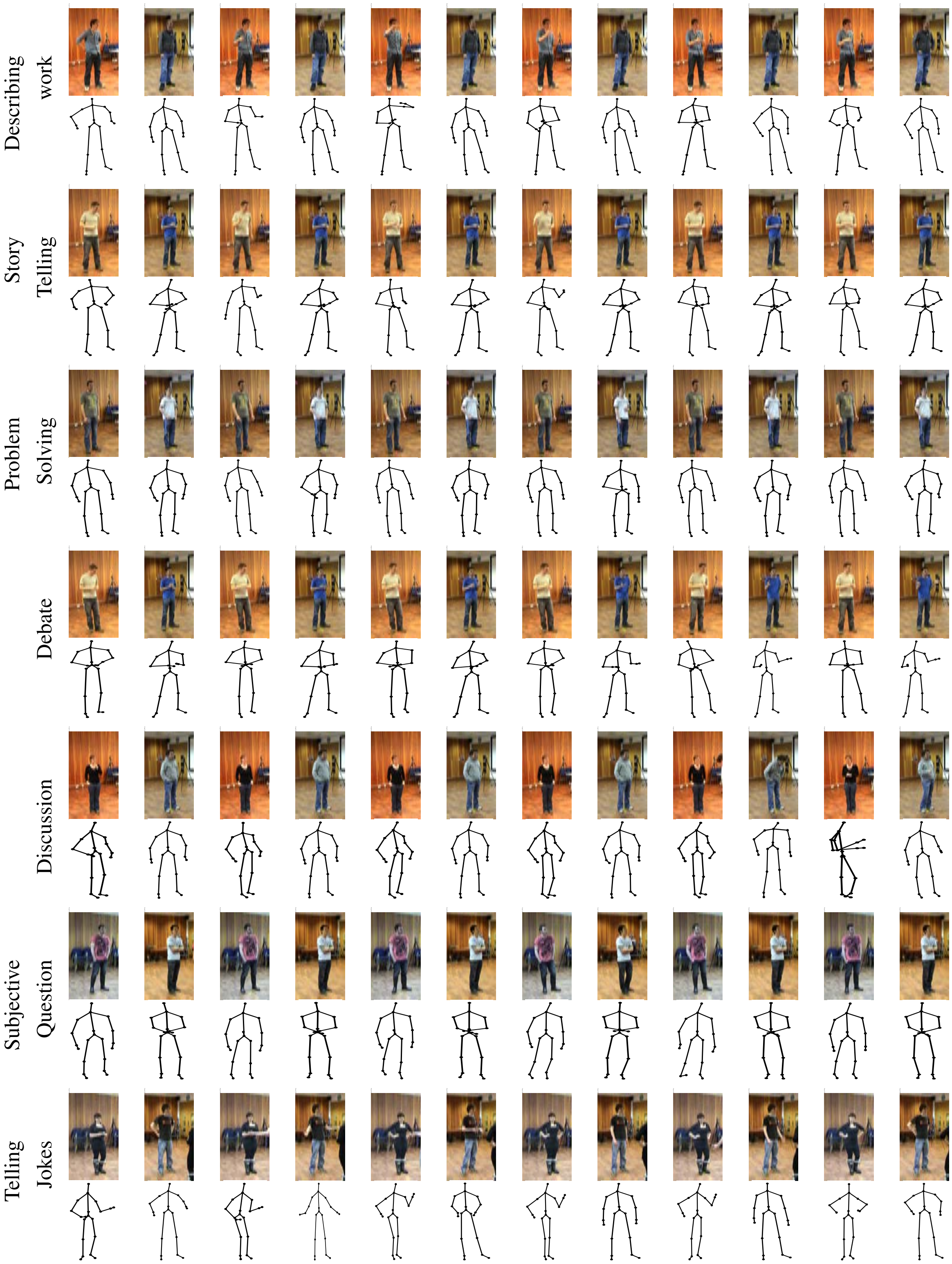


Figure 1. Example images and 3D skeletons from 7 different scenarios. The time difference between each consecutive frame shown is two seconds. Example videos of seven different scenarios are available online[1]. Note that the RGB images were captured by separately synchronized cameras at different viewing angles to Kinect - hence the discrepancy in pose. The RGB data is not used in this study.

ing, 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, i.e. to identify the content of a conversation using pose features only.

The rest of the paper is organized as follows. Section 2 gives details of data acquisition. The proposed method is presented in Section 3, which includes low level feature extraction, feature generalization and classification. Experimental results and discussions are in Section 4. Section 5 concludes the paper.

2. Data acquisition

Action recognition systems can often be built on relatively easy to extract low level features such as temporal SIFT features [20] or temporal Harris corner features [13]. Typically, those actions can be easily distinguishable from a visual perception point view, e.g. waving, jumping, and punching. The dataset used for training and evaluation can thus be labeled using those action types. More subtle behaviors, such as grooming, drinking and eating, can also be distinguished [6, 11]. These primitive action and short time span behavior can be well defined, semantically. Thus, the data can be labeled to individual, relatively short sequences. However, social interactions are more complex and difficult to recognize since the actions, motions and motivations of multiple people must be understood. Each of those interactions can contain multiple types of primitive actions. Often, it is the temporal dynamics of those primitive motions, actions and interactions that differentiate one from another. For example, two people having a debate may have very similar primitive motions and actions to having a discussion a topic , although the event as whole can be considered different in the context of conversational interaction. Thus, it is unrealistic to label each and every primitive action in the sequences of conversational interactions since the sequences are usually thousands of times longer. It is also not necessary as those primitive action labeling alone doe not describe the whole event. Hence, in this study we directly use the conversational topic or the nature of the conversation to label the whole sequence and pose the question that whether it is possible to distinguish different types of conversation using 3D gesture alone. The conversational categories are subtly different to each other, which poses a great challenge for recognition.

In this work, we choose seven categories and use a two-Kinect set-up to record 3D human pose. Each person was recorded using a Kinect Sensor, which captured pose at 30fps. Each of the cameras was slightly 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 from opposing view points on an issue we gave to them. In the fifth task they were asked to discuss the issues surrounding a particular statement and come to agreement whether they believe the statement is true or not. The participants were asked to trying to reach an agreement through discussion; hence, it is different to the debate task. 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.

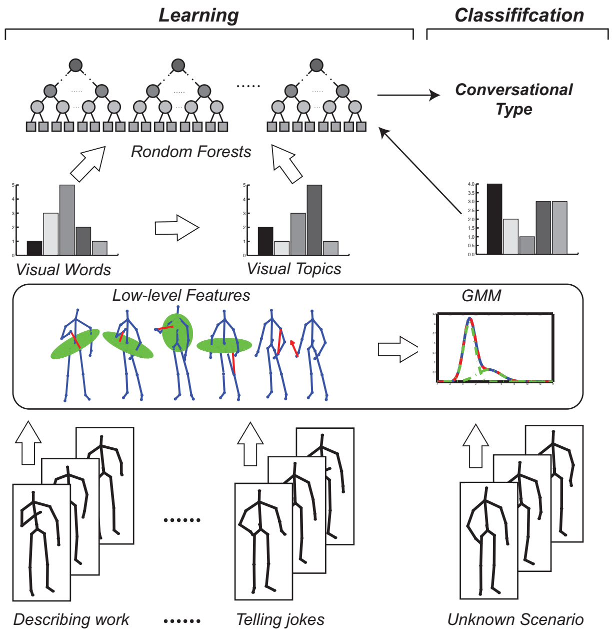


Figure 2. Flowchart of the proposed method.

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 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 1. The full dataset used in this study is available for download from this address http://csvision.swan.ac.uk/converse.html.

3. Proposed method

The proposed method first extract displacement and velocity measurements from the Kinect output. Wavelet de-

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.

composition is then applied to extract low level features from each of those measurements. The wavelet coefficients represent sudden changes in measurements at different temporal scales, and they are treated as the low level motion features. A temporal generalization of those features are then carried out to encapsulate temporal dynamics, which first produces a visual vocabulary and then further generalized them to visual topics through Latent Dirichet Allocation analysis. A discriminative model based on Random Forests is then trained and applied to classify different types of conversational interactions. The flowchart shown in Fig. 2 illustrates the steps from pose measurements, to wavelet analysis, to unsupervised clustering and generalization, and to supervised classification.

3.1. Low level feature extraction

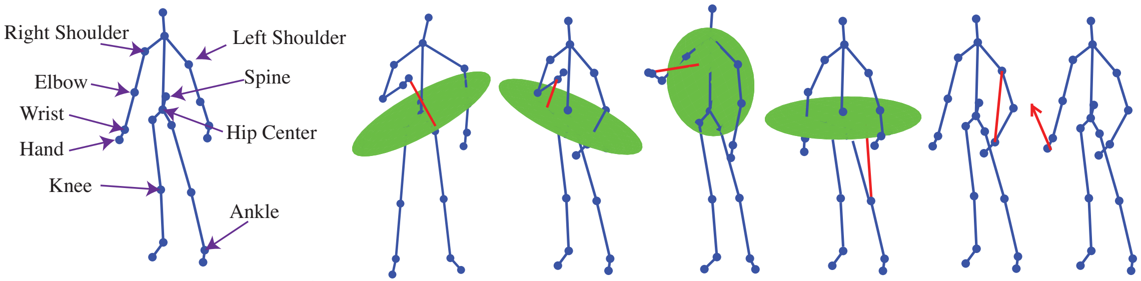
3D poses have been shown to be useful in motion capture data retrieval and action recognition. Motivated by existed work, such as [2, 12, 17], we extract three types of pose measurements to depict the pose and motion of the body. These geometry measurements extracted from a kinematic chain are simple but useful for representing human gesture and motion over time. These measurements are then decomposed to wavelet coefficients and treated as low level features. Briefly, the first set of measurements are the distance between two joints at different time intervals and is depicted in Fig. 3(f). The second set measures the distance between a joint and reference planes defined using different parts of the body (see Fig. 3(b,c,d,e)). The third set measures the velocity of individual joints (see Fig. 3(g)).

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. 3(b)), and the other is defined by the vector connecting the spine and right shoulder (Fig. 3(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. 3(d)), which is used to measure movements of lower arm joints from both arms. The overlapping in measurement is to make sure that the 3D motion of those joints are captured among those 2D measurement combinations. The fourth plan is perpendicular to the third plan and crossing the same spine point (Fig. 3(e)). This reference plan is used to measure movement of knees and ankles (ankle points are more stable than feet in Kinect estimation). Next, we provide the definition for each measurement of joint movement.

The 3D location of a joint at time 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 are defined 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 measurement $\digamma_d$ representing the Euclidean distance between joints over Δt is defined as: $\digamma_d = D\{(\omega_{i,t}), (\omega_{j,t+\Delta t})\}$. If $i = j$, then the it measures the distance of movement of the joint over time Δt, otherwise, it measures the distance between two different joints separated by time.

The measurements $\digamma_{pd1}$ and $\digamma_{pd2}$ are the shortest distance from joint $\omega_{n,t}$ to the plane $\phi_{ijk,t+\Delta t}$



(a) Kinect Skeleton Model (b) (c) (d) (e) (f) (g)

Figure 3. Visualization of the pose measurements. (b) - (e) The distance of a joint to a reference plane. (f) Illustrates the distance between joints feature. (g) The joint velocity

and the plane $\psi_{ijk,t+\Delta t}$, respectively. They 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 joint velocity along the direction of the vector $\pi_{ij,t+\Delta t}$ and vector $\pi_{ijk,t+\Delta t}$, respectively. They are defined as: $F_{jv} = V\{(\omega_{n,t}), (\pi_{ij,t+\Delta t})\}$ and $F_{pv} = V\{(\omega_{n,t}), (\pi_{ijk,t+\Delta t})\}$

Thus, 42 different low-level pose measurements are extracted from the Kinect data, with $\Delta t = 1.0s$, by computing the displacement distances, velocity of both left and right limbs are computed according to these four reference planes.Table 2 summarizes different types of measurements. It is notable that we selected 34 measurements from upper body joints, and 8 measurements from lower body joints.

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. In this work, we apply wavelet decomposition to emphasize sudden changes in those measurements at multiple scales. Wavelet analysis has been widely used in signal processing, e.g. texture analysis [19], due to its ability to analyze signal in spatial - spatial frequency domain. Here, we consider the changes of the low level relative motion in local temporal region can be used as clues for conversational scenario classification. The strength of the motion in the short time window is represented by the coefficients. For simplicity and in the interest of keeping the feature dimension space lower, we adopt the Daubechies 2 wavelet (Haar), whose mother wavelet function is defined as

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2} \\ -1 & \frac{1}{2} \leq t < 1 \\ 0 & otherwise \end{cases} \tag{1}$$

and scaling function is defined as

$$\phi(t) = \begin{cases} 1 & 0 \leq t < 1 \\ 0 & otherwise \end{cases} \tag{2}$$

Fig. 4 illustrates an example of wavelet decomposition, from which we may see that abrupt changes in measurement are highlights in the wavelet coefficients across the scales. In total, 29 scales are used for each measurement. That is there are forty two 29-dimensional feature spaces.

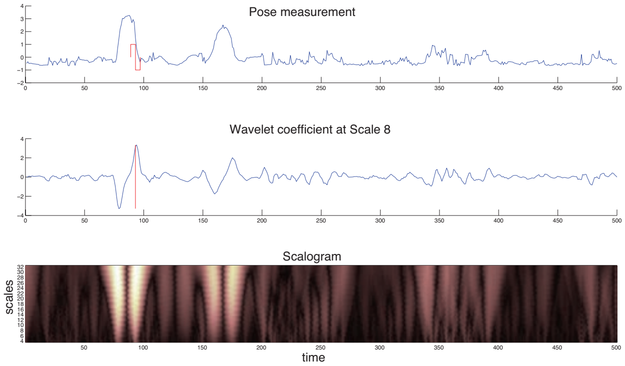


Figure 4. An example of decomposing one of the temporal measurements using Daubechies 2 wavelets.

3.2. Dynamic feature descriptors

3.2.1 Visual words

The extracted low level pose features are direct measurements of relative motion at a short time window. In order to capture the dynamics in interaction, we generalize those low level features to a middle level to summarize the distributions of those primitive motions in a reasonable time span, i.e. 500 frames or 20 seconds in our case. Furthermore, since we are classifying conversational scenarios at 20-second segments, the common approach of appending feature vectors will result in prohibitively long feature vectors for discriminative classifiers to train. In this work, we thus adopt the bag of words approach to derive middle level features that are suitable for classification of conversational interactions, each of which may contain various amount of primitive motions. 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, for example, the locations of joints in the kinematic model. Con-

Table 2. Pose motion measurements. (b), (c), (d) and (e) denote the reference planes as shown in Fig. 3.

Joint	Reference Plane or Joint	Type	Number of measurements
hands, wrists, and elbows at $t + \Delta t$	hands, wrists, and elbows at t	displacement	6
hands, and wrists at $t + \Delta t$	shoulders at t	displacement	4
hands, wrists, and elbows at $t + \Delta t$	reference planes (b & c) at t	displacement	6
hands, wrists, and elbows at $t + \Delta t$	reference plane (b & c) at t	velocity	6
hands, wrists, and elbows at $t + \Delta t$	reference planes (d) at t	displacement	6
hands, wrists, and elbows at $t + \Delta t$	reference planes (d) at t	velocity	6
knees, and ankles at $t + \Delta t$	reference plane (e) at t	displacement	4
knees, and ankles at $t + \Delta t$	reference plane (e) at t	velocity	4

sequently, we are concerned with the distributions of those features across time. We hence use 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 actions.

As a result of low level feature analysis, there are forty two 29-dimensional features spaces, each of which corresponds to one measurement from Kinect sensor. To generalize visual words in each of the 42 feature space, we apply the Gaussian Mixture Model (GMM), which is a common and powerful method in parameterizing complex, often multi-modal distributions. It approximates the underlying distribution using a number of Gaussian components. A GMM can be formulated as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \qquad (3)$$

where K is the number of the components, π_k is the mixing coefficients and $\mathcal{N}$ denotes the normal distribution with mean μ_k and covariance Σ_k. The mixing coefficients π_k must satisfy the constrains $\sum_{k=1}^{K} \pi_k = 1$ and $0 < \pi_k < 1$. These components $\mathcal{N}(x|\mu_k, \Sigma_k)$ are combined with different weighting π_k to provide a multi-modal density.

Given wavelet coefficients $X = \{x_1, x_2, ..., x_n, ...x_N\}$, temporally collected into each 29-dimensional feature space, the parameters of the GMM, π, μ and Σ are estimated by maximizing the log likelihood function given by:

$$ln\ p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\} \qquad (4)$$

The EM algorithm is the most popular algorithm for finding maximum likelihood solution to Equation 4.

For each feature space, one GMM model is fitted across whole data set, and the Gaussian clusters are used to form the a visual vocabulary. Each GMM component is considered as a visual word. A further generalization of these visual words can be carried out based on the distribution of visual words in an extended time span that is often larger than typical primitive action.

3.2.2 Visual topics

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.* [18] applied the LDA model to extract action categories from low-level spatial-temporal words in an unsupervised fashion. Inspired this work, we use LDA to generalize the learned visual words to form visual topics that are learned across feature spaces, instead of individual feature spaces as in the case for visual words.

We assume that those learned visual words are generated by a mixture of visual topics. To learn those visual topics, we split the sequences into 500 frames (20 seconds) sections each of which is considered as a visual document that contains multiple visual topics. The LDA model 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.

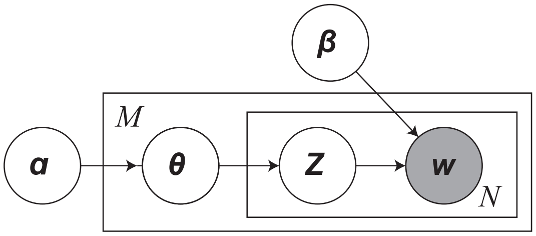


Figure 5. Latent Dirichlet Allocation (LDA) model

Briefly, the LDA model illustrated in Figure 5 was firstly proposed by David Blei *et. al.* [3] in 2003, which is similar to Probabilistic Latent Semantic Analysis (pLSA) [9], but with assumption of having a Dirichlet prior. In the LDA model, the outer plate represents the replicated documents (in our case, 20 seconds clips), and the inner plate represents the repeated topics and words. It is notable that the parameters α and β is corpus-level parameters which determine the mixing proportions of the topics $\{\theta_{d=1}...\theta_{d=M}\}$, and the Dirichlet prior on the per topic-word distribution respectively, where M is the number of documents. The parameters θ_d are the document-level parameter, which are generated once per document. In each document, the word-level parameters Z_n and W_n are sampled once per word.

In our case, given the model α, β the visual words W can be generalize in following way:

1. The number of visual words is determined by Poisson process: $N \sim Poisson(\xi)$;
2. The mixture proposition of visual topics θ_d is chosen according to Dirichlet process: $\theta_d \sim Dir(\alpha)$;
3. For each of the N words W_n:
   (a) Firstly, a visual topics is chosen by multinomial process: $Z_n \sim Multinomial(\theta_d)$;
   (b) Secondly, a visual word is generated according to $p(W_n|Z_n, \beta)$, a multinomial probability with condition on the visual topics Z_n.

Given a corpus, a set of visual documents with a number of visual words, the latent visual topic for each visual word can be obtained by applying Bayesian inference. The joint distribution of a topic θ, a set of N visual words generated according to a set of N visual topics is given by:

$$p(\theta, Z, W|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(Z_n|\theta)\, p(W_n|Z_n, \beta) \quad (5)$$

The marginal distribution of a visual document can be computed by integrating over θ and summing over Z:

$$p(W|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{Z_n} p(Z_n|\theta)\, p(W_n|Z_n, \beta) \right) d\theta \quad (6)$$

Thus, given a visual words, the posterior probability of its latent visual topic can be inferred according to Bayesian theory, as follows:

$$p(\theta, Z|W, \alpha, \beta) = \frac{p(\theta, Z, W|\alpha, \beta)}{p(W|\alpha, \beta)} \quad (7)$$

Approximation inference methods such as variational inference [3], Gibbs sampling [8], and expectation propagation [15] may be adopted to efficiently solve (7).

Next, we use the distributions of those visual words and topics to classify different conversational scenarios.

3.3. Classification

A discriminative classifier, namely Random Forests [4] is employed in this work, to evaluate the discriminative power of our features, and to investigate whether classifying conversational scenarios is possible by merely using 3D pose features.

Random Forests (RF) illustrated in Figure 6 is an ensemble classifier consisting of a set of decision trees, which significantly improves the generalization ability of the classifier compared to a single decision tree. At the bootstrap aggregating stage (bagging), assuming that the data sample is independent and identically distributed, new training sets are generated by randomly sampling with replacement from the complete training set. For each new training set, one decision tree is constructed which consists of a set of split nodes and linking edges. Each non-leaf node stores a random test function which is applied to the input data, and leads to the leaf node. In the leaf nodes, the final predictor is stored. At the prediction stage, all the trees classify the incoming data independently, the most voted class given by the trees is considered as the final classification of the forest. This is illustrated in 6.

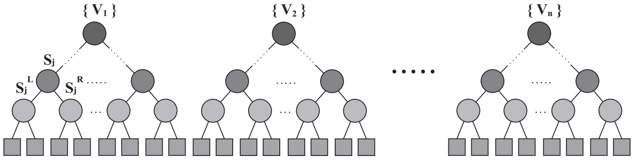


Figure 6. Random Forests

To train and test the classifiers, each recorded sequence was split into 500 frames sections. Each section was labeled as the task from which it was extracted and used as a single example, both for training and testing. As described in Section 3.2.1, both visual words and visual topic were extracted. In order to investigate the discriminative ability of this two types of features, we train the RF classifiers on these features separately to compare the recognition result. Given a set of sections with class labels, a histogram of visual words and visual topics are obtained for each section. The parameters of Random Forests is learned based on those histograms. We learn 100 decision trees for Random Forests by randomly sampling with replacement from the complete training set.

4. Experimental results

The human conversational interaction dataset was collected following the approach described in Section 2, and used in the presented experiments. All tasks were completed by 8 different pairs of people in 482 minutes, which resulted in 869,142 frames in total. The full dataset is available for download from this address http://csvision.swan.ac.uk/converse.html. 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.

The 3D pose measurements were exacted directly from the Kinect output. Wavelet decomposition was then ap-

Table 3. Average classification results using visual features from only single participant.

	Visual words		Visual topics	
	k-NN	RF	k-NN	RF
Describing Work	67.4	71.5	61.7	65.9
Story Telling	47.5	59.2	47.5	59.2
Problem Solving	55.9	66.6	47.5	65.5
Debate	42.7	69.9	62.4	69.0
Discussion	49.1	53.9	49.2	53.9
Subjective Question	47.2	75.7	53.6	66.0
Telling jokes	23.9	63.6	28.1	57.9
Average	47.7	65.8	50.0	62.5

Table 4. Average classification results using visual features from paired participants.

	Visual words		Visual topics	
	k-NN	RF	k-NN	RF
Describing Work	92.4	92.4	92.4	92.4
Story Telling	72.3	72.3	72.3	72.3
Problem Solving	42.5	60.0	42.5	62.5
Debate	88.0	100.0	82.7	91.4
Discussion	82.0	82.0	82.0	82.0
Subjective Question	60.0	80.0	70.0	80.0
Telling jokes	50.0	90.0	60.0	90.0
Average	69.6	82.3	71.7	81.5

plied to individual measurement and each produced a 29-dimensional feature space, 29 wavelet scales, as a low level representation. As the length of sequences across different tasks and subjects is different, in order to avoid the bias, the GMMs were fitted to the features that were sampled from these sequence with equal number of samples. Each feature space produced 10 visual words, and there were 42 features spaces in total. The Kinect sequences are then labeled by those visual words. These sequences were partitioned into segments of 20 seconds long, where the visual words were collected and form a visual document for each segment. A total of 25 visual topics from 420 visual words were inferred by LDA model using Gibbs sampling method. The histogram of visual words and visual topics for each 20 seconds segment was then computed, and used as higher level feature descriptors. To carry out the classification, 10-fold 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. This is necessary to avoid over-fitting. In addition to the Random Forests classifier, K-nearest neighbor (K-NN) classifier with $k = 5$ was also used. Both classifiers were trained on the same training set independently.

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 3 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 47.7% and 65.8% were achieved by K-NN and RF classifiers, respectively. The Random Forests classifier clearly outperformed K-NN. When using visual topics, which produces significantly shorter feature vectors (25 vs 340), there was slight decrease in the advantage of using Random Forests. However, RF's superior performance was still statistically significant. The further generalization from visual words to visual topic did not deliver real difference in terms of accuracy. This could be explained as that the visual topic may be a good generalization and interpolation of the motion and gesture of conversational interactions but it slightly sacrificed its discriminative power. These results, however, are very interesting, as they suggest that there is a good chance to distinguish different types of conversation merely by observing gestures of a single person from the pair.

For the next experiment we combine features from two participants by concatenating their features before feeding into the classifiers. The results are summarized in Table 4. There were broad improvements reported by all both classifiers. The confusion matrix given by the Random Forests classifier using visual words descriptor is shown in Table 5. The averages are 69.6% and 82.3% by K-NN and Random Forests, respectively. It is worth noting that the true positive rates for scenarios, "problem solving", "subjective question" and "telling jokes" reported by K-NN were 42.5%, 60.0% and 50.0%, compared to 60.0%, 80.0% and 90% given by the Random Forests, which suggests the conversational types cannot be successfully classified merely by the nearest neighbor in feature space. For visual topics, the length of each descriptor is 50 which is far more less compared with the visual words descriptor, 840. However, as shown in Table 6 and the Random Forests confusion matrix in Table 6, similar results still can be achieved, which means the discriminative power of visual topics is still acceptable after temporal generalization. The significant overall performance increase compared to using feature from single participant clearly highlights the benefit of having multiple streams of information when observing people during an interaction.

The results we have achieved suggested that it is feasible and practical to discriminatingly classify conversational interactions just based on human poses alone. 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

Table 5. Confusion matrices by Random Forests classification using visual words.

		Work	Story	Problem	Debate	Discussion	Question	Joke
Random Forests	Work	92.4	0.0	0.0	0.0	7.6	0.0	0.0
	Story	0.0	72.3	0.0	10.0	0.0	0.0	17.6
	Problem	5.0	07.5	60.0	0.0	20.0	0.0	7.5
	Debate	0.0	0.0	0.0	100	0.0	0.0	0.0
	Discussion	0.0	0.0	0.0	8.0	82.0	0.0	10.0
	Question	0.0	0.0	0.0	10.0	10.0	80.0	0.0
	Joke	0.0	0.0	0.0	10.0	0.0	0.0	90.0

Average = 82.3

Table 6. Confusion matrices by Random Forests classification using visual topics.

		Work	Story	Problem	Debate	Discussion	Question	Joke
Random Forests	Work	92.4	0.0	0.0	0.0	7.6	0.0	0.0
	Story	17.6	72.3	0.0	0.0	0.0	10.0	0.0
	Problem	0.0	0.0	62.5	07.5	0.0	7.5	22.5
	Debate	8.5	0.0	0.0	91.4	0.0	0.0	0.0
	Discussion	0.0	10.0	8.0	0.0	82.0	0.0	0.0
	Question	10.0	0.0	0.0	10.0	0.0	80.0	0.0
	Joke	0.0	0.0	0.0	10.0	0.0	0.0	90.0

Average = 81.5

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. Random Forests was applied to classify different types of conversational interactions. The results also suggest 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. An even larger data set and perhaps more sophisticated modeling techniques should be investigated as future work. However, we believe this work offer a somewhat different perspective to action and interaction analysis.

References

[1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Survey*, 43(3):16, 2011. 1

[2] G. F. A.Yao, J. Gall and L. V. Gool. Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11. BMVA Press, 2011. 1, 4, 5

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003. 6, 7

[4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 7

[5] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu. Poseshop: Human image database construction and personalized content synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 19(5):824–837, 2013. 1

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE international workshop on: Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005. 3

[7] A. Fathi. Social interactions: A first-person perspective. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 1226–1233, Washington, DC, USA, 2012. IEEE Computer Society. 1

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004. 7

[9] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999. 6

[10] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multiview videos: Comparative explorations of recent developments. *Selected Topics in Signal Processing, IEEE Journal of*, 6(5):538 –552, sept. 2012. 1

[11] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang. Action detection in complex scenes with spatial and temporal am-

biguities. In *IEEE International Conference on Computer Vision*, 2009. 3

[12] L. Kovar and M. Gleicher. Automated extraction and parameterization of motions in large data sets. *ACM Trans. Graph.*, 23(3):559–568, Aug. 2004. 4

[13] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005. 3

[14] D. McColl, Z. Zhang, and G. Nejat. Human body pose interpretation and classification for social human-robot interaction. *International Journal of Social Robotics*, 3(3):313–332, 2011. 1

[15] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc., 2002. 7

[16] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006. 1

[17] M. Müller, T. Röder, and M. Clausen. Efficient content-based retrieval of motion capture data. *ACM Trans. Graph.*, 24(3):677–685, July 2005. 4

[18] J. 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, 2008. 6

[19] T. Randen and J. H. Husoy. Filtering for texture classification: A comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291–310, 1999. 5

[20] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 357–360, New York, NY, USA, 2007. ACM. 3

[21] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. Motion capture from body-mounted cameras. *ACM Trans. Graph.*, 30(4):31:1–31:10, July 2011. 1

[22] V. Singh and R. Nevatia. Simultaneous tracking and action recognition for single actor human actions. *The Visual Computer*, 27(12):1115–1123, 2011. 1

[23] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transaction on Circuits Systems for Video Technology*, 18(11):1473–1488, 2008. 1