Conversational Interaction Recognition Based on Bodily and Facial Movement

Jingjing Deng, Xianghua Xie(✉), and Shangming Zhou

Department of Computer Science, Swansea University, Swansea, UK
x.xie@swansea.ac.uk
http://csvision.swan.ac.uk

Abstract. We examine whether 3D pose and face features can be used to both learn and recognize different conversational interactions. We believe this to be among the first work devoted to this subject and show that this task is indeed possible with a promising degree of accuracy using both features derived from pose and face. To extract 3D pose we use the Kinect Sensor, and we use a combined local and global model to extract face features from normal RGB cameras. We show that whilst both of these features are contaminated with noises. They can still be used to effectively train classifiers. The differences in interaction among different scenarios in our data set are extremely subtle. Both generative and discriminative methods are investigated, and a subject specific supervised learning approach is employed to classify the testing sequences to seven different conversational scenarios.

Keywords: Human interaction modeling · Conversantional interaction analysis · 3D human pose · Face analysis · Randomized decision trees · HMM · SVM

1 Introduction

There has been some success in using features extracted from high-level information such as body pose, e.g. automatically learning sign language to perform classificaiton task [5]. However, assumptions about the subjects in the scenes, such as body orientation, are routinely made to constrain the solution. A further problem with studying social interaction is that there are often occlusion since usually participants would face one another, meaning observations are often incomplete. For this reason, often the interactions examined are less intimate and can be viewed at a coarser resolution. For example Zhang *et al.* [21] studied group interactions in a work meeting between multiple people, detecting events such as presenting to the group, conducting a group discussion or note taking etc. This is achieved by first estimating the state of each participant and then using this information to infer the group action. Decomposing the group interaction

A. Campilho and M. Kamel (Eds.): ICIAR 2014, Part I, LNCS 8814, pp. 237–245, 2014.
DOI: 10.1007/978-3-319-11758-4_26

into a two level process of firstly inferring what each person is doing, and then from this deducing the group action is a common approach [1,19,21]. Probabilistic models such as Hidden Markov Models (HMM) can be employed to overcome noisy observations, both at the image level and on the person dependent action classification level. However, for this approach to be effective there needs to be an understanding of which motions, poses or gestures that an individual performs is likely to be an important building block. Often this is dependent on the granularity of the actions being observed.

In order to understand the high-level semantic human activity, accurate pose estimation is generally required. To perform such as task using RGB cameras, e.g. [8,9], remains an open challenge. In [10,11], we proposed to leverage recent advances in technology in extracting 3D pose using a consumer sensor (Microsoft Kinect) to examine the feasibility of recognizing human interactions between two people using the body pose only. Rather than recognizing just key social events, we attempt to analyze and classify different conversational interactions. In this work, we investigate both bodily and facial pose features for recognizing the type of conversation they are conducting. We do not examine strongly differentiable interactions, such as high-tempered arguments or disputes, as in previous research efforts studying interaction. Neither do we employ the use of actors. Different from affect recognition, where a single observation can typically be used to identify the affective state (e.g. smile implies happiness), there is not a direct connection between a single observation and the type of the conversation being performed; rather it is the sequence of observations as an interaction is in progress and is of importance. We acknowledge that bodily and facial movements are not necessarily generalizable across subjects. Here, we aim to find out whether it is possible to generalize subject specific motion cues which can be used to identify the topic of a conversation.

2 Data Acquisition

Data was collected using a multi-camera set-up. Each person was recorded using a Kinect Sensor, which captured pose at 30fps. The face images were captured using two high definition cameras operating at 25fps. The first task was to discuss an area of current work that the participant was undertaking. 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 tell jokes in turn. In total, there are about 8 hours long Kinect sequences and equal length of face sequences. The dataset is available for download from http://csvision.swan.ac.uk/converse.html.

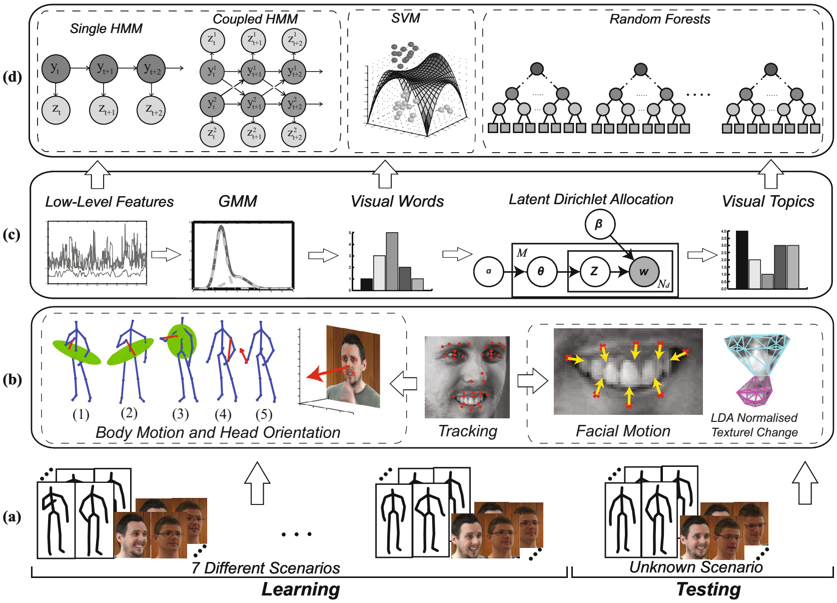

Fig. 1. Flowchart of the proposed method

3 Methodology

The proposed method first extract motion features from Kinect output and localize facial fiducial points in RGB face images using a two level shape model. The head orientation is then computed based on face localization and is treated as part of the pose feature. The localization of face fiducial points also provides two sets of features: shape and appearance. The shape features are derived from the coefficients of a global shape model that is used for face localization. The appearance features are obtained from the textural coefficients of two local face models after Linear Discriminant Analysis. Hidden Markov Models (HMMs) are then used to model the conversational interactions based on these low level features at individual time instances. Interactions between pair of subjects are captured using coupled HMM. A temporal generalization of both pose and face features are also carried out to encapsulate temporal dynamics, which first produces a visual vocabulary features and then further generalizes them to visual topics through Latent Dirichlet Allocation analysis. Discriminative classifiers, Support Vector Machine (SVM) and Random Forests (RF), are applied to classify interactions into seven different scenarios. Moreover, we apply modulator functions to those mid level features so that we can learn the importance of those individual features, which is then used in the SVM classification. Fig. 1 illustrates the steps from low level feature extraction, to unsupervised feature generalization, and to supervised modeling and classification.

3.1 Pose Feature Extraction

Motivated by recent work, such as [2,16,17], we extract three types of low level 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 measures the distance between two joints at different time intervals. The second set of feature measures the distance between a joint and a reference plane defined using different parts of the body. The third set of feature measures the velocity of individual joints. These are depicted in row (b) in Fig. 1.

In this study, we use three reference planes, (1) (2) and (3) showed in row (b) in Fig. 1. The first two reference planes, (1) and (2) 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. 1, row (b), (1)), and the other is defined by the vector connecting the spine and right shoulder (Fig. 1, row (b), (2)). 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. 1, row (b), (3)), 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.

3.2 Face Feature Extraction

The face images acquired have varied poses and sometimes contain occlusions (e.g. glasses and hand movement). Consequently, holistic models, such as active appearance models, [6], have been found not robust enough to track the faces beyond a few dozens of frames. We thus integrate the local component shape models with a global shape model [12]. We use the point distribution model [6] to build two local shape models, which are trained using feature points from upper and lower faces, respectively, with overlapping nostril fiducial points. The two models hence are focusing on local deformations at eyes and mouth regions that are important to model interactions. The overlap provides a weak constraint between two local models. The result from local models provides a good initialization for the second level global shape model. Each of the fitness function is composed of a texture cost and a shape cost. Response scores based on Haar-like rectangular features [20] and the GentleBoost algorithm [13] are used to evaluate the texture fitness. We follow [7] to formulate a generic shape cost function, which is applied to both local and global models. The two level fitness functions are then optimized using the simplex algorithm.

Based on the localization results, two types of features are extracted to capture facial dynamics: shape and appearance. For each face image there are 35 fiducial points, many of which are for localization purposes, and are not contributing to deformations. We hence project those localized points to the global shape model space learned at the localization stage and retain 90% eigenvalue, which results in

9-dimensional shape features. This dimensionality reduction is also desirable for training classifiers. For appearance feature, we similarly project the facial texture to a PCA texture model that is learned from the training samples used for localization. Since there are significant differences between the upper part and lower part of the face, two separate PCA models are built. Again, 90% eigenvalue is retained, which results in 14-dimensional features for both upper part and lower part. However, for appearance feature we also perform a Linear Discriminant Analysis [3] to minimize the individual textural characteristics in derived appearance features. We re-project the coefficients back into the texture subspace and calculate the residue, which is used as the final appearance feature. Thus, a total of 37-dimensional features are learned for capturing facial dynamics.

3.3 Head Orientation Estimation

Currently the Kinect sensor has the ability of facial tracking and head pose estimation. However, the performance and accuracy are greatly affected by the data acquisition environment and experiment set-up, especially the imaging distance and the participant's pose. Hence, we perform head orientation estimation by extending the results from face tracking. As part of facial feature extraction, we obtain a set of five fiducial points for each face image: two external eye corners, two mouth corners, and nose tip. We follow the work by Gee and Cipolla [14] to estimate the head orientation from a single image using these fiducial points.

3.4 Temporal Feature Descriptors

To determine which conversational scenario directly based on short-term, primitive actions is unlikely going to be successful. Instead, the temporal dynamics of those short-term motions and primitive actions are useful in revealing the topic of conversation. To capture such dynamics, we employ Hidden Markov Model (HMM) which is well suited to model temporal sequential data. However, we also attempt to generalize those face and pose features to a middle level to summarize the distributions of those primitive motions in a reasonable time span, 5 seconds in our case. The common approach of appending feature vectors will result in prohibitively long feature vectors for discriminative classifiers to train. 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.

The Latent Dirichlet Allocation (LDA) model [4] has been widely used to discover abstract "topics" from a collection of words or low level features, e.g. [18]. In this work, we 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.

We first construct a visual vocabulary by fitting Gaussian Mixture Model to each dimension of the low-level 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 20 seconds sections, each of which is considered as a visual document that contains multiple visual topics. The LDA model is learned by using Gibbs sampling inference method, [15], and applied to extract interaction categories from low level temporal visual words. The distribution of both visual words and visual topics are used as temporal feature descriptors for conversational interaction modeling and classification.

3.5 Modeling Using Coupled HMM

In order to explicitly model the dependence between the two subjects we use separate HMM to represent each person and then adding an edge between the subjects across time to build a Coupled HMM (CHMM), e.g. [19]. Row (d) in Fig. 1 depicts the CHMM used in this work. To perform classification, CHMMs are learned for each of the seven classes, $\{\Lambda_1, .., \Lambda_7\}$. Given a set of T observations $Z_T = \{z_T, z_{T-1}, .., z_1\}$ from an unknown class we classify it to the model that maximizes $p(\Lambda_n|Z_T)$, where n denotes class ID. This is calculated in two stages. Firstly the forward-backward algorithm is used to calculate $p(Z_T|\Lambda_n)$ by recursively computing $p(y_t = j|z_{t-1}, .., z_1, \Lambda_n) = \sum_{i=1}^{m} \mathbf{A}_{ij} p(z_{t-1}|y_{t-1} = i) p(y_{t-1} = i|z_{t-2}, .., z_1, \Lambda_n)$, where $\mathbf{A}$ denotes the transition matrix, and then summing the probabilities over all states in the final time instance, i.e. $p(Z_T|\Lambda_n) = \sum_{i=1}^{m} p(z_T|y_T = i) p(y_T = i|z_{T-1}, .., z_1, \Lambda_n)$, following which, $p(\Lambda_n|Z_T)$ can be calculated using Bayes' rule assuming a flat prior across all classes.

3.6 Classification Using Discriminative Classifiers

Whilst generative models, such as HMM, is important in explaining the data, discriminative ones tend to be more effective in classification tasks. In this work, we also employ SVM and Random Forests to study the discriminative power of the features, and only the middle level features are used since a concatenation of low level features will result in a too large dimensional feature space.

3.7 Classification Using SVM Ranked Features

In order to automatically identify the influential features from high dimensional space, we conduct feature ranking via a scheme that applies the entropy regularization and particle swarm optimization (PSO) techniques to the construction of an optimal SVM model [22]. The novelty of this scheme lies in that the model selection, feature identification and dimensionality reduction are performed simultaneously in an integrated manner. During learning process the importance of less influential attributes automatically approaches to zero, whilst the importance of the most important attributes turns to be one. As a result, only the most influential features remain in the final SVM model.

Specifically, given a data set $\{x_l, y_l\}_{l=1}^{N_p}$used for performing model selection by the PSO, where $y_l \in \{-1,\ 1\}$ denotes the label of data x_l and N_p denotes the number of classes, the following fitness function is used to identify optimal hyper-parameters for SVM: $f = \frac{1}{N_p} \sum_{k=1}^{N_p} \left(\bar{y}_k - y_k\right)^2 + \lambda_1 \left(- \sum_{i=1}^{n} \theta_i \log(\theta_i)\right) + \lambda_2 \left(\sum_{i=1}^{n} \theta_i\right)$, where $\theta_i \in (0,\ 1)$ indicates the importance of the input variable to the classification task, λ_i (> 0) are called regularization coefficients, $\bar{y}_k$ are the labels predicted by the SVM model. The second term, an entropy penalty, is used to remove redundant features. Because the entropy distribution of importance ranks would become zero (minimum) if importance values of features reach $\{0, 1\}$, during the training process the importance ranking values associated with redundant features would be forced to approach to zero and the importance ranks associated with influential features would move towards one. The third term encourages feature sets that are as compact as possible.

4 Results and Discussions

All 7 tasks were completed by 8 different pairs of people in a total of 482 mins, producing a total of 869,142 pose frames and 724,285 RGB face images. Together with estimated head orientation, 35 low level pose features were extracted. 37 low level face features were derived from face localization. To extract the visual words, for each feature, a Gaussian Mixture Model with 10 components was fitted to the low level features across different pairs. In order to extract the visual topic from the visual word, the sequences were chopped into 20-second sections, each of which was considered as a visual document. We learned LDA models with 25 visual topics for pose and face separately, and each visual word was inferred and assigned with a potential visual topic. Finally, at the scenario classification stage, each recorded sequence is split into 5-second sections. For the discriminative classifiers, the histogram of visual words or topics is computed, and used as a feature vector for each section. For the CHMM, the feature vector of every 10 frames, for the sake of computational feasibility, in the section corresponds to an observation node expanded across time. To carry out the classification, 10-fold cross validation is adopted. Note, neighboring segments are not distributed across different folders.

The results of using CHMM are summarized as following. Using face and pose features alone achieved 53.2% and 55.9% respectively, compared to a random

Table 1. Classification results using visual words (%)

	Face&Pose			
	KNN	RF	SVM	SVM-R
Describing Work	81.2	90.6	88.4	100.0
Story Telling	59.7	51.0	70.6	80.2
Problem Solving	41.4	12.8	35.1	80.7
Debate	55.3	51.6	67.7	91.8
Discussion	50.0	62.7	69.5	61.1
Subjective Question	30.8	5.2	35.8	91.7
Jokes	36.3	14.2	47.7	80.0
Average	50.7	41.2	59.3	89.1

Table 2. Classification results using visual topics (%)

	Face&Pose			
	KNN	RF	SVM	SVM-R
Describing Work	63.5	91.7	76.4	100.0
Story Telling	35.1	73.2	68.3	80.2
Problem Solving	37.1	73.6	74.3	80.7
Debate	48.6	73.6	67.1	81.97
Discussion	38.4	78.7	63.5	61.11
Subjective Question	22.5	63.3	63.5	91.74
Jokes	27.5	70.3	66.3	80.0
Average	38.9	74.9	68.5	87.3

chance of around 14%. The combination of face and pose feature achieved an average of 59.6%. When using visual words and visual topics, the performance decreased significantly. With visual words, overall accuracy of 32.0%, 33.6% and 36.4% were produced using face, pose, face and pose, respectively. After further generalization to visual topic, its performance reduced further to 28.3%, 30.8% and 30.7%. This was generally expected, since the feature generalization causes an enhancement of commonality among different scenarios, which caused HMMs modeling slightly more common features and hence reduced their discriminative power.

Next, we tested the mid level features with discriminative classifiers, i.e. SVM and RF, see Tables 1 and 2. The classification results are considerably better. For example, the overall accuracy using standard SVM with face and pose visual words achieved 59.3%, compared with a mere 36.4% achieved by CHMM. With visual topics, the difference is even more evident: 68.5% vs. 30.7%. The combination of pose and face features showed markable improvements over using face or pose features alone. We also present the results using KNN. With visual words, RF was inferior to others and SVM is clearly performed better. With further generalized features, there are clear improvements for both RF and SVM, but not for KNN, and RF slightly out-performed SVM.

However, using our SVM ranked features, there were substantial improvements for all features and raised the performance close to 90%. It is evidently clear that feature selection is important in differentiating different conversation scenarios.

Whilst the Kinect sensor permits direct estimation of 3D pose that is currently more robust and accurate than RGB camera methods, the accuracy of the data collected still contains some noise, as does the face features used in this work. However, despite this we have shown that good recognition of conversational interactions can still be achieved. The suggests that it is possible to recognize the conversational topics based on gesture and facial dynamics.

References

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

2. Yao, A., Gall, J., Fanelli, G., Gool, L.V.: Does human action recognition benefit from pose estimation? In: BMVC (2011)
3. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs fisherfaces: recognition using class specific linear projection. IEEE T-PAMI **19**(7), 711–720 (1997)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. of Machine Learning Research **3**, 993–1022 (2003)
5. Buehler, P., Everingham, M., Zisserman, A.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR (2009)
6. Cootes, T., Edward, G., Taylor, C.: Active appearance models. IEEE T-PAMI **23**(6), 681–685 (2001)
7. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. PR **41**, 3054–3067 (2008)
8. Daubney, B., Xie, X.: Entropy driven hierarchical search for 3d human pose estimation. In: BMVC, pp. 1–11 (2011)
9. Daubney, B., Xie, X.: Tracking 3d human pose with large root node uncertainty. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1321–1328 (June 2011)
10. Deng, J., Xie, X., Daubney, B.: A bag of words approach to subject specific 3d human pose interaction classification with random decision forests. Graphical Models **76**(3), 162–171 (2014)
11. Deng, J., Xie, X., Daubney, B., Fang, H., Grant, P.W.: Recognizing conversational interaction based on 3D human pose. In: Blanc-Talon, J., Kasinski, A., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2013. LNCS, vol. 8192, pp. 138–149. Springer, Heidelberg (2013)
12. Fang, H., Deng, J., Xie, X., Grant, P.: From clamped local shape models to global shape model. In: IEEE ICIP, pp. 3513–3517 (September 2013)
13. Friedman, J., Hastie, T., Tibshirani, R.: Addictive logistic regression: a statistical view of boosting. Annals of Statistics **28**, 337–407 (2000)
14. Gee, A.H., Cipolla, R.: Determining the gaze of faces in images. IVC **12**, 639–647 (1994)
15. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America **101**, 5228–5235 (2004)
16. Kovar, L., Gleicher, M.: Automated extraction and parameterization of motions in large data sets. ACM ToG **23**(3), 559–568 (2004)
17. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. ACM ToG **24**(3), 677–685 (2005)
18. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. IJCV **79**(3), 299–318 (2008)
19. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. IEEE T-PAMI **22**(8), 831–843 (2000)
20. Viola, P., Jones, M.: Robust real-time face detection. IJCV **57**(2), 137–154 (2004)
21. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I.: Modeling individual and group actions in meetings with layered hmms. IEEE Multimedia **8**(3), 509–520 (2006)
22. Zhou, S.M., Lyons, R.A., Bodger, O., Demmler, J.C., Atkinson, M.A.: Svm with entropy regularization and particle swarm optimization for identifying childrens health and socioeconomic determinants of education attainments using linked datasets. In: IEEE Inter. Conf. Neural Networks, pp. 3867–3874 (2010)