HIDDEN CONDITIONAL RANDOM FIELDS FOR MEETING SEGMENTATION

Stephan Reiter, Björn Schuller and Gerhard Rigoll

Technische Universität München
Institute for Human-Machine Communication
Arcisstraße 21, 80290 München, Germany
{reiter, schuller, rigoll}@tum.de

ABSTRACT

Automatic segmentation and classification of recorded meetings provides a basis towards understanding the content of a meeting. It enables effective browsing and querying in a meeting archive. Though robustness of existing approaches is often not reliable enough. We therefore strive to improve on this task by applying conditional random fields augmented by hidden states. These Hidden Conditional Random Fields have been proven to be efficient in low level pattern recognition tasks. Now we propose to use these novel models to segment a pre-recorded meeting into meeting events. Since they can also be seen as an extension to Hidden Markov Models an elaborate comparison of the two approaches is provided. Extensive test runs on the public M4 Scripted Meeting Corpus prove the great performance of applying our suggested novel approach compared to other similar methods.

1. INTRODUCTION

Automatic analysis of meetings is of growing interest in the research community and beyond. The potential possibility of creating meeting minutes automatically, and the countless applications that are about to evolve from electronic data processing of meetings make this topic highly exciting.

Segmenting a pre-recorded video of a meeting is a task that is performed by few research groups only. The approach that is most commonly applied are Hidden-Markov-Models (HMM) [1], an efficient generative model with a hidden state-structure. This approach is used by McCowen et al [2] with quite reasonable results. More sophisticated methods based on HMM have also been developed, resulting in asynchronous multi-stream HMM or coupled HMM [2]. In addition to that approaches for the segmentation of meetings in two successive steps have been developed, either using two consecutive HMM [3] or using two different probabilistic classifiers like Multi-layer-perceptrons [4] or Long-Short-Term-Memory [5] to improve the segmentation results. However all these generative models assume that the observations are conditionally independent. This restriction is sometimes too strict, especially when there are long-range dependencies between observations.

Conditional Random Fields (CRF), first introduced by Lafferty et al [6], use an exponential distribution to model a sequence given the observation sequence. This avoids the independence assumption between observations, and allows non-local dependencies between state and observation. Additionally CRF allow unnormalized transition probabilities. Furthermore a Markov assumption can still be enforced allowing the inference to be performed efficiently by using dynamic programming. CRF have been applied in various tasks as part-of-speech tagging and information extraction [6].

CRF assign a label for each observation (each frame of a time-sequence) and do not directly provide a way to estimate the conditional probability of a class label for an entire sequence. Therefore a generalized CRF with hidden state sequences is used, so called Hidden Conditional Random Fields (HCRF). This kind of model was introduced by Quattoni [7] and Gunawardana [8] and successfully applied for Gesture Recognition [9] and Phone Classification [8]. HCRF are able to deal with features that can be arbitrary functions of the observations without complicating the training [8] but in this work we intend to compare the performance of HCRF to standard Maximum Likelihood trained HMM.

We apply trained HCRF to find a optimal segmentation of a meeting, where the segments have a length of several seconds. The detected segments are called meeting events and describe group actions as discussion, monologue or presentation.

The remainder of the paper is organized as follows: Section 2 describes the database we used. In Section 3 the used features are described. Section 4 then gives an overview of the applied models and in Section 5 the results are presented.

2. MEETING CORPUS

Within our research we use the publicly available M4 Scripted Meeting Corpus, described in [10]. It consists of fully scripted meetings recorded in a Smart Meeting Room at IDIAP, equipped with fully synchronized multichannel audio and video recording facilities. Each of the recorded par-

Meeting Event	Train	Test
Discussion	48	49
Monologue 1	14	12
Monologue 2	10	13
Monologue 3	10	14
Monologue 4	9	10
Note-taking	6	3
Presentation	11	18
White-board	16	20
Total	124	139

Table 1. Number of meeting events in training and test sets

ticipants had a close-talk lapel microphone attached to his clothes. An additional microphone array was mounted on top of one center meeting table. Video signals were recorded onto separate digital video tape recorders by three television video cameras, providing PAL quality.

Each captured meeting consists of a set of predefined group actions in a fixed order defined in an according agenda. The appearing group actions are:

- Discussion (all participants engage in a discussion)
- Monologue (one participant speaks continuously without interruption)
- Note-taking (all participants write notes)
- Presentation (one participant at front of the room presents using the only projector screen)
- White-board (one participant at front of the room talks and makes notes on the white board)

In each meeting there were four participants at six possible positions: four seats plus white-board and presentation board. The number of different meeting events in the different data sets is summarized in table 1.

The database comprises a total of 59 scripted meetings with two disjoint sets of participants. A fixed training set makes use of 30 videos, while the remaining 29 are used throughout evaluation.

3. MULTI-MODAL FEATURE EXTRACTION

For each participant person-specific features were extracted from the cameras, the lapel microphones, and the microphone array. Therefore we make use of visual as well as audio features. The person-specific video features are:

- head vertical centroid
- head eccentricity
- right hand horizontal centroid
- right hand angle
- right hand eccentricity
- head and hand motion

For each video frame areas of skin color are detected by a Gaussian mixture model. Next the greatest skin color blobs are identified as face using a face detector and described by the vertical centroid and eccentricity. From the remaining blobs the one with the rightmost horizontal position is regarded as hand and is represented by its horizontal position, eccentricity, and angle. For more detail on the video features please refer to [3]

In addition to the visual features we also used person-specific audio features extracted from the lapel microphones and the microphone-array:

- speech activity from each seat
- speech relative pitch
- speech energy
- speech rate

As speech activity measure SRP-PHAT was used. Pitch was extracted using a SIFT algorithm and normalized to the mean value. All used features and methods to derive them are explained in more detail in [3].

In addition to the individual multi-modal features group features were extracted from the white-board and projector-screen area. In detail they include the speech activity from the white-board and projector screen as audio features. From the visual information the mean difference between a current frame and a reference background image is used.

4. HIDDEN CONDITIONAL RANDOM FIELDS

An Hidden Conditional Random Field (HCRF) models the conditional probability of a class label k given the observation sequence $\mathbf{o} = (o_1, o_2, \ldots, o_T)$:

$$p(k|\mathbf{o}, \lambda) = \frac{1}{z(\mathbf{o}, \lambda)} \sum_{\mathbf{s} \in k} e^{\lambda \cdot f(k, \mathbf{s}, \mathbf{o})} \quad (1)$$

Hereby λ denotes the parameter vector and f is the vector of sufficient statistics as it is called with Conditional Random Fields (CRF). $\mathbf{s} = (s_1, s_2, \ldots, s_T)$ is a hidden state sequence that is passed through during the calculation of the conditional probability. The function $z(\mathbf{o}, \lambda)$ ensures that the model forms a properly normalized probability and is defined as

$$z(\mathbf{o}, \lambda) = \sum_{k} \sum_{\mathbf{s} \in k} e^{\lambda \cdot f(k, \mathbf{s}, \mathbf{o})} \quad (2)$$

The choice of the vector $f(k, \mathbf{s}, \mathbf{o})$ determines the probability that can be modeled by the HCRF. In [8] it is shown that choosing the vector f in the right manner results in an HCRF which is equivalent to a conventional HMM model. As the HCRF do not obey the strict rules of normalization the transitions do not necessarily need to sum to one, neither do the observations be proper probability densities. However in this work the topology of the HCRF is restricted to obey a Markov

chain. So it is possible to apply efficient algorithms, known from HMM for example, to calculate the probability given in equation (1). Also a direct comparison of the performance of HCRF and HMM becomes feasable. As the structure of the HCRF in this work is bound to such conditions, the conditional probability of HCRF can be rewritten in a form known from HMM, by defining transition scores a_{ij} and observation scores $b_i(o_t)$:

$$\begin{aligned} a_{ij} &\widehat{=} e^{\lambda_{i,j}^{(Tr)}} \\ b_i(o_t) &\widehat{=} e^{\lambda_i^{(occ)}+\lambda_i^{(M1)}o_t+\lambda_i^{(M2)}o_t^2} \end{aligned} \tag{3}$$

Analogous to HMM the conditional probability of a model can then efficiently be calculated using the forward and backward recursions using the transitions and observation scores of equation (3). The forward variables $\alpha_t(i)$ are given by

$$\begin{aligned} \alpha_{t+1}(j) &= \left(\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right)b_j(o_{t+1}) \\ &= \left(\sum_{i=1}^{N}\alpha_t(i)e^{\lambda_{i,j}^{(Tr)}}\right)e^{\lambda_i^{(occ)}+\lambda_i^{(M1)}o_{t+1}+\lambda_i^{(M2)}o_{t+1}^2} \end{aligned} \tag{4}$$

where N is the number of hidden states of the model.

The backward recursions $\beta_t(j)$ can be estimated using the familiar recursion:

$$\begin{aligned} \beta_t(i) &= \sum_{j=1}^{N}a_{ij}b_j(o_{t+1})\beta_{t+1}(j) \\ &= \sum_{j=1}^{N}e^{\lambda_{i,j}^{(Tr)}}e^{\lambda_i^{(occ)}+\lambda_i^{(M1)}o_{t+1}+\lambda_i^{(M2)}o_{t+1}^2}\beta_{t+1}(j) \end{aligned} \tag{5}$$

Using the forward variables $\alpha_t(i)$ the probability $p(\mathbf{o}|k,\lambda)$ of a model of class k generating the observation $\mathbf{o}$ can now be written as:

$$p(\mathbf{o}|k,\lambda) = \sum_{i=1}^{N}\alpha_T(i) \tag{6}$$

and therefore the conditional probability of a class label k given the observation becomes

$$p(k|\mathbf{o},\lambda) = \frac{\sum_{i=1}^{N}\alpha_T(i)}{\sum_k\sum_{i=1}^{N}\alpha_T(i)} \tag{7}$$

Defining the HCRF in this way allows to use dynamic programming techniques a Forward-Backward and Viterbi for decoding as with HMM. It can further be shown [8] that by setting the parameters λ in the following way an HCRF gives the conditional probability density as an HMM with transition probabilities a_{ij}, emission means μ_i, and emission covariances σ_i observing a observation of D dimensions:

$$\begin{aligned} \lambda_{ij}^{(Tr)} &= \log a_{ij} \\ \lambda_i^{(Occ)} &= -\frac{1}{2}\left\{\log\left[(2\pi)^D\prod_{d=1}^{D}\sigma_{i,d}^2\right]+\sum_{d=1}^{D}\frac{\mu_{i,d}^2}{\sigma_{i,d}^2}\right\} \\ \lambda_{i,d}^{(M1)} &= \frac{\mu_{i,d}}{\sigma_{i,d}^2} \\ \lambda_{i,d}^{(M2)} &= -\frac{1}{2}\frac{1}{\sigma_{i,d}^2} \end{aligned} \tag{8}$$

Hereby i and j denote various states of the models, whereas d denotes one dimension. For simplicity reasons to avoid aberrant indices, the above equations only considers only one mixture component. But it is easy to extend this expression to multi-mixture models as they are commonly used.

Using a HCRF initialized according to equation (8) a direct and fair comparison of the two different models can be performed.

5. EXPERIMENTS

In our presented work we apply the HCRF on the task of meeting event segmentation and recognition. To guarantee equal conditions we train an HMM on the training set (cp. section 2) using standard Maximum Likelihood training using the Baum-Welch algorithm [1]. Then we build a HCRF with the same number of states and "mixtures", and determine the parameters λ using equation (8). Then Viterbi decoding of both models is performed using exactly the same procedure on the test set. The performance of the tested models is measured by an established accuracy measure, known from the speech recognition community. This measurement is defined as the sum of insertions (Ins), deletions (Del), and substitutions (Sub), divided by the total number of events in the groud truth defined by manually labelling the meeting corpus:

$$Accuracy = \left(1-\frac{Ins+Del+Sub}{TotalEvents}\right)\times 100\% \tag{9}$$

Some selected results of our extensive experiments are presented for various numbers of states and mixtures in table 2. It is remarkable that HCRF significantly outperform HMM of exactly the same structure in almost all cases. The mean difference between the performance of HCRF compared to HMM is 3.31% absolute. The standard deviation is 2.20. The maximum gain of HCRF is 10.07% using 20 states and 2 mixtures. Only in three of 170 tested cases there was a slight degradation of the accuracy. Using a HCRF-model with five states and one gaussian gives the best overall result with an accuracy of 92.09%. Hereby there are only two substitutions (white-board is regarded as presentation twice), eight

States	Mixtures	HMM	HCRF
4	4	88.49%	89.93%
5	1	88.49%	**92.09%**
6	1	87.77%	89.93%
9	1	83.45%	91.37%
19	1	87.05%	89.93%
20	1	84.89%	89.21%
20	2	61.15%	71.22%

Table 2. Results of the comparative approach of HMM and HCRF

deletions, and one insertion. This outcome is a result that bears a high potential in it, since the HCRF are not trained at all after the transfer of the parameters. So we expect the recognition rates to raise even further when a training as suggested in [11] is performed.

6. CONCLUSION

HCRF are known to be a powerful tool to use in recognition and segmentation tasks like phone classification and gesture recognition. In this work we proved the high capability of this model in another pattern recognition task, the segmentation of meeting events. In comparison to HMM with the same underlying structure HCRF proved to be highly effective and outperformed the appropriate HMM clearly. The 92.09% accuracy yielded by HCRF is one of the best results ever reported on this task on this database. Further improvements are expected by training the HCRF after initializing the parameter from an equivalent HMM.

7. ACKNOWLEDGMENTS

This work is supported by the European IST Programme Project FP6-0033812. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

8. REFERENCES

[1] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[2] Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Dong Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 305–317, March 2005.

[3] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, Iain McCowan, and Guillaume Lathoud, "Modeling Individual and Group Actions in Meetings With Layered HMMs," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, 2006.

[4] Stephan Reiter, Björn Schuller, and Gerhard Rigoll, "Segmentation and recognition of meeting events using a two-layered HMM and a combined MLP-HMM approach," in *Proceedings of the International Conference on Multimedia and Expo (ICME)*, Toronto, Ontario, Canada, July 2006.

[5] Stephan Reiter, Björn Schuller, and Gerhard Rigoll, "A combined LSTM-RNN - HMM - approach for meeting event segmentation and recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[6] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning (ICML)*, Massachusetts, USA, June 2001, pp. 282–289, Morgan Kaufmann, San Francisco, CA.

[7] Ariadna Quattoni, Michael Collins, and Trevor Darrell, "Conditional random fields for object recognition," in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., pp. 1097–1104. MIT Press, Cambridge, MA, 2005.

[8] Asela Gunawardana, Milind Mahajan, Alex Acero, and John C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, Lisbon, Portugal, September 2005.

[9] Sy Bor Wang, Ariadna Quattoni, Louis-Philippe Morency, David Demirdjian, and Trevor Darrell, "Hidden conditional random fields for gesture recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[10] Iain McCowan, Samy Bengio, Daniel Gatica-Perez, Guillaume Lathoud, Florent Monay, Darren Moore, Pierre Wellner, and Herve Bourlard, "Modeling human interaction in meetings," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003.

[11] Milind Mahajan, Asela Gunawardana, and Alex Acero, "Training algorithms for hidden conditional random fields," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.