EXPLORING CONTEXTUAL INFORMATION IN A LAYERED FRAMEWORK FOR GROUP ACTION RECOGNITION

Dong Zhang, Samy Bengio

IDIAP Research Institute, Martigny, Switzerland
Swiss Federal Institute of Technology, Lausanne, Switzerland
{zhang, bengio}@idiap.ch

ABSTRACT

Contextual information is important for sequence modeling. Hidden Markov Models (HMMs) and extensions, which have been widely used for sequence modeling, make simplifying, often unrealistic assumptions on the conditional independence of observations given the class labels, thus cannot accommodate overlapping features or long-term contextual information. In this paper, we introduce a principled layered framework with three implementation methods that take into account contextual information (as available in the whole or part of the sequence). The first two methods are based on state *alpha* and *gamma* posteriors (as usually referred to in the HMM formalism). The third method is based on Conditional Random Fields (CRFs), a conditional model that relaxes the independent assumption on the observations required by HMMs for computational tractability. We illustrate our methods with the application of recognizing group actions in meetings. Experiments and comparison with standard HMM baseline showed the validity of the proposed approach.

1. INTRODUCTION

Most of the existing work on sequence modeling has used Hidden Markov Models (HMMs) [7] and extensions, including coupled HMMs, input-output HMMs, multi-stream HMMs, and asynchronous HMMs (see [5] for a recent review of models). However, HMM-based approach has one well-noted weakness: the assumption on the conditional independence of observations given the class labels. Therefore, complex features, such as overlapping and neighboring features, which take into account the long-term contextual information, cannot be used in HMM-based approaches.

However, it is widely known that contextual information is important for sequential activity recognition. For instance, it may be hard to predict the current activity state solely based on past activities and current observation. A more superior method of classification should incorporate a broader series of consecutive observations both before and after the current time in consideration. Such contextual information is essential for sequence modeling.

A multi-layer framework was introduced in [10] for group action recognition in meetings. The fundamental idea is that, by defining an adequate set of individual actions, we can decompose the group action recognition problem into two levels, from individual to group actions. The output of individual action layer provides the input to group action layer.

The focus of this paper is to present three methods of implementing such a layered framework that can take into account the contextual information. The first two methods are based on state *alpha* and *gamma* posterior definitions (as usually referred to in the HMM formalism). The state *alpha* and *gamma* posterior can take into account the context information since it is defined as the probability of being in a state given the part or the whole observation sequence (see Section 2.2 for details). The third method is based on Conditional Random Fields (CRFs), a conditional model that relaxes the independent assumption on the observations required by HMMs for computational tractability. A key advantage of CRFs is their great flexibility to include a wide variety of arbitrary, non-independent features of the input. Thus, CRFs can also take into account the contextual information (see Section 2.3 for details). All the three methods bring improvement over the standard HMMs method, which reflects on the results obtained on a 59-meeting corpus, for a set of eight group actions.

The paper is organized as follows. Section 2 introduces the multi-layer framework and the three implementation methods. Experiments and discussion are presented in Section 3. Conclusions are drawn in Section 4.

2. THE MULTI-LAYER FRAMEWORK

The layered framework is illustrated in Figure 1. Details on the framework have been reported in [10]. In the next sections, we first briefly describe the layered framework for group action recognition in meetings, followed by introducing the ideas of "alpha", "gamma" and CRFs.

2.1. Framework Overview

Let *I-HMM* denote the lower recognition layer for individual action, and *G-HMM* denote the upper layer for group action. *I-HMM* receives as input audio-visual features extracted from each participant, and outputs recognition results, in the form of posterior probabilities (α or γ as defined in the following

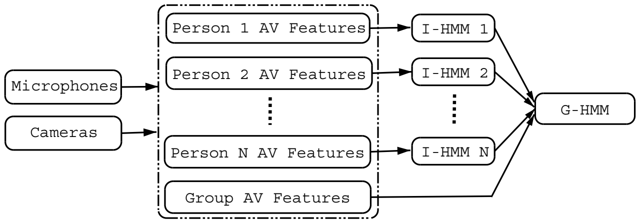

Fig. 1. The multi-layer framework applied to group action recognition: the lower layer recognizes individual actions of participants using low-level audio-visual (AV) features. The output of this layer provides the input to the second layer, which models interactions. Individual actions naturally constitute the link between the low-level audio-visual features and high-level group actions.

sections). In turn, *G-HMM* receives as input the output from *I-HMM*, and a set of *group features*, directly extracted from the raw streams, which are not associated to any particular individual. In our framework, each layer is trained independently, and can be substituted by any of the HMM variants that might capture better the characteristics of the data.

We next present three implementations of such a layered framework. To facilitate description, we first define the following symbols:

- the *whole* observation: $X = x_1^T = \{x_1, ..., x_t, ..., x_T\}$
- the *past* observation: $x_1^t = \{x_1, x_2, ..., x_t\}$
- the *future* observation: $x_{t+1}^T = \{x_{t+1}, x_{t+2}, ..., x_T\}$
- the observation within the window: $x_{t-c}^{t+c} = \{x_{t-c}, x_{t-c+1}, ..., x_t, ..., x_{t+c-1}, x_{t+c}\}$
- q_t: the HMM state at time t.

In Baum-Welch algorithm [1], also known as the Forward-Backward procedure, we define,

- Forward variable $\alpha(i,t) \stackrel{\text{def}}{=} P(x_1^t, q_t = i)$: the probability of having generated the sequence x_1^t and being in state i at time t.
- Backward variable $\beta(i,t) \stackrel{\text{def}}{=} P(x_{t+1}^T | q_t = i)$: the probability to generate the rest of the sequence x_{t+1}^T given that we are in state i at time t.
- Variable $\gamma(i,t) \stackrel{\text{def}}{=} P(q_t = i | x_1^T)$: the probability being in state i at time t given the observation sequence x_1^T.

2.2. "Alpha" and "Gamma"

In this method, we first train HMMs for the individual action layer. The output of the individual action layer is in the form of α-based features. The output feature vector serves as the input to the upper layer, which is also trained independently. This method takes into account of the contextual information based on the "past" observation , *i.e*, the observation sequence upon current time t: $x^t sequence_1$. The linked features are defined as the probability of state i given x_1^t is $P(q_t = i|x_1^t)$,

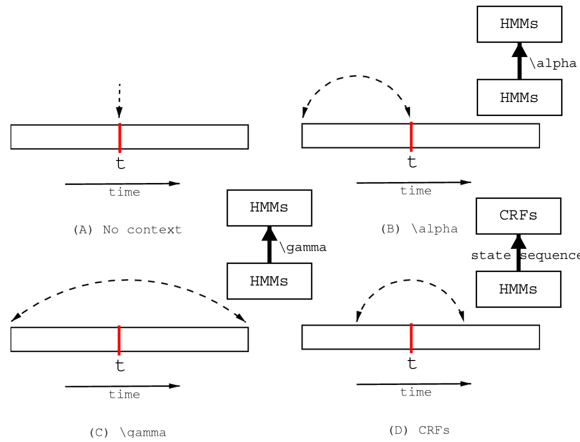

Fig. 2. Illustration of context information (Note that t indicates the current time): (A) without contextual information; (B) "Alpha" taking into account the contextual information of the *past* observation from the beginning upto the current time X_1^t; (C) "Gamma" taking into account the contextual information of the *whole* observation X_1^T; (D) Conditional random field taking into account the arbitrary contextual information.

which can be calculated as:

$$P(q_t = i|x_1^t) = \frac{P(q_t = i, x_1^t)}{P(x_1^t)} = \frac{\alpha(i,t)}{\sum_{j=1}^{N_S} \alpha(j,t),} \tag{1}$$

where N_S is the total number of states. Obviously, $P(q_t = i|x_1^t)$ is a posterior probability measure so that

$$\sum_{i=1}^{N_S} P(q_t = i|x_1^t) = 1 \tag{2}$$

The second method based on γ is similar to the above method based on α. In this method, we take into the contextual information of the *whole* observation sequence (both past and the future sequence): x_1^T. The probability of state i given the whole sequence x_1^T is defined as γ. The variable $\gamma(i,t)$ can be expressed in terms of the forward-backward variables,

$$\gamma(i,t) = \frac{\alpha(i,t)\beta(i,t)}{\sum_{i=1}^{N_S} \alpha(i,t)\beta(i,t)}, \tag{3}$$

where N_S is the total number of states. Note that the normalization factor $\sum_{i=1}^{N_S} \alpha(i,t)\beta(i,t)$ makes $\gamma(i,t)$ a posterior probability measure so that

$$\sum_{i=1}^{N_S} \gamma(i,t) = 1 \tag{4}$$

2.3. Conditional Random Fields

Conditional random fields (CRFs), a special case of undirected graphical model shown in Figure 3, were introduced originally by [4] for modeling sequences. Recently, there has been an explosion of interest in CRFs, with successful applications including text processing [6, 9], bio-informatics [8], and computer vision [3].

The underlying idea of CRFs is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The primary advantage of CRFs over hidden Markov models is their conditional nature, resulting in the relaxation of the independence assumptions required by HMMs in order to ensure tractable inference. Therefore, CRFs has the great flexibility to include a wide variety of arbitrary, non-independent features of the input. As illustrated in Figure 2 (D), we can see that CRFs can take into account the contextual information by defining arbitrary, non-independent features.

Now let $X = \{x_1, x_2, ...x_T\}$ be the observed input data sequence. Let Y be a set of states, each of which is associated with a label and $\{y_1, y_2, ...y_T\}$ is a sequence of states. Linear-chain CRFs thus define the conditional probability of a state sequence given an input sequence to be

$$P(Y|X) = \frac{1}{Z_o} \exp(\sum_{t=1}^{T} F_\theta(y_t, y_{t-1}, X)), \quad (5)$$

where Z_o is a normalization factor over all state sequences. CRF is in terms of exponentiated feature functions F_θ, computed in terms of weighted sums over the features of the cliques. In particular,

$$\sum_{t=1}^{T} F_\theta(y_t, y_{t-1}, X) = \sum_j \lambda_j t_j(y_{t-1}, y_t, X, t) + \sum_k \mu_k s_k(y_t, X, t)), \quad (6)$$

where $t_j(y_{t-1}, y_t, x, t)$ is a transition feature function of the entire observation sequence and the labels at positions t and $t-1$ in the label sequence; $s_k(y_t, x, t)$ is a state feature function of the label at position t and the observation sequence; and λ_j and μ_k are parameters to be estimated from training data. The CRFs training and decoding can be performed using gradient descent and Viterbi algorithms (for more details, please refer to [9]).

3. EXPERIMENTS

3.1. Data Sets, Actions, and Audio-Visual Features

We use a set of 59 five-minute, four-participant meetings, recorded in a room equipped with three cameras and 12 microphones. Although the meetings were recorded according to a script for turn-taking patterns, the participants behavior was unconstrained and reasonably natural. A sets of individual action (like *writing*, *speaking*) and group actions (like *discussion*, *monologue*, or *presentation*) have been defined.

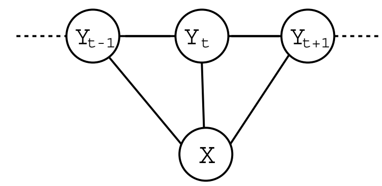

Fig. 3. Conditional Random Fields: the hidden nodes can depend on observations at any time step, thus relaxing the independence assumptions required by HMMs.

The monologue action is further distinguished by the person actually holding the monologue (e.g. monologue 1 is meeting participant one speaking). We also define combinations of two parallel actions (like a presentation and note-taking). The investigated actions are multimodal, we therefore we extracted a set of generic features, including audio features derived from microphone arrays and lapel microphones, and visual features extracted from skin color blobs from each participant. Details on data sets, action lexicons and audio-visual features have been reported in [10].

3.2. Results and Discussions

We investigated the following four configurations:

- **Single-layer HMMs:** A normal HMM is trained using the audio-visual features extracted from each participant and concatenated together.
- **Alpha:** Using the α posterior probability outputs from individual action layer as input to the group action recognition layer (Section 2.2).
- **Gamma:** Using the γ posterior probability outputs from individual action layer as input to the group action recognition layer (Section 2.2).
- **HMM + CRFs:** We use HMM modeling individual action. The output is a sequence of a state sequence resulted from Viterbi decoding. The output state sequence serves as input to the group action layer, which is modeled using conditional random fields (Section 2.3).

The data set is divided into 30 meetings for training, and 29 for testing. For training, we used ten-fold cross-validation to select the hyper-parameters (*i.e.* the number of states, the number of Gaussian). After the best parameters were chosen, we re-trained models on the whole training set and applied the models on the test set. The results are summarized in Table 1, Figure 4 and Figure 5, in terms of *Frame Error Rate* (FER) and *Action Error Rate* (AER) respectively.

We first discuss results in terms of FER shown in Figure 4. We can observe that (1) the multi-layered methods (using α, γ, or CRFs) always out-perform the single-layer HMMs. For example, the α-based multi-layer approach produced 19% FER, which is 6% absolute improvement over using the single-layer HMMs. This improvement is statistically significant with a confidence level above 95%, using a

Table 1. Results of four methods in terms of FER and AER.

Method		FER (%)	AER (%)
Single-layer HMMs		25.42	23.47
Multi-layer	α	19.32	16.83
	γ	17.47	15.42
	CRFs	13.56	15.11

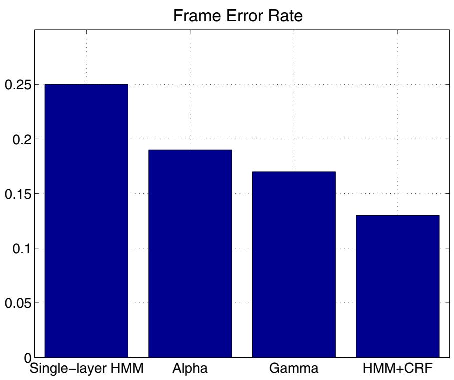

Fig. 4. Comparison of the four methods – the single-layer HMMs, the multi-layer approach based on α, γ, and CRFs – in terms of FER (*frame error rate*).

standard proportion test [2]. (2) Regarding the three multi-layer methods, CRFs produced the best results with the FER of 13%, which is 4% absolute improvement over the second best method based on γ significant at 95% confidence level. There might be two reasons. First, CRFs take into account contextual information by including a wide variety of arbitrary, non-independent features of the observation sequence. Thus CRFs are more flexible than α and γ. Second, CRFs are a conditional model training for maximizing the posterior probability over label sequences given the observation sequence, rather than a joint distribution over both label and observation sequences in HMMs. (3) We can also see that γ outperforms α. This is not surprising given that γ takes into account the *whole* observation sequence as contextual information while α only takes into account the *part* observation contextual information upon the the current time.

In terms of AER, we can observe the same trend. The multi-layer approaches always outperform the single-layer HMMs statistically significant at 95% confidence level. CRFs got the best performance among the three methods, although the improvements are not statistically significant given the few number of group actions.

4. CONCLUSIONS

We addressed the problem of recognizing group actions in meetings with a layered framework. We presented three implementation methods (*alpha*, *gamma*, CRFs) that can take

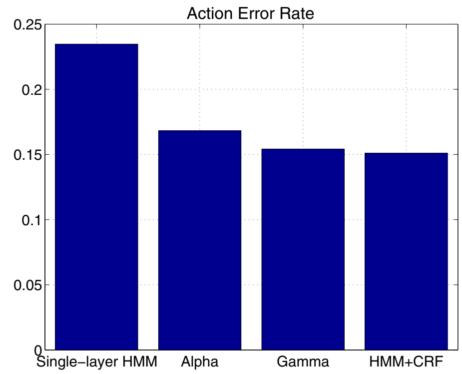

Fig. 5. Comparison of the four methods – the single-layer HMMs, the multi-layer approach based on α, γ, and CRFs – in terms of AER (*action error rate*).

into account the contextual information. The state alpha and gamma can take into account the context information by definitions as the probability of being in a state given the *part* or *whole* observations. CRFs takes into account contextual information by including a wide variety of arbitrary, non-independent features. Experiments on a public 59-meeting corpus demonstrate the effectiveness of the proposed methods to recognize a set of eight group actions.

Acknowledgments

This work is supported by the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (IM2) and European IST Programme Project FP6-0033812.

5. REFERENCES

[1] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. In *ICSI-TR-97-021*, 1997.

[2] D. Gibbon, R. Moore, and R. Winski. Handbook of standards and resources for spoken language systems. In *Mouton de Gruyter*, 1997.

[3] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. In *In Neural Information Processing Systems 16. MIT Press, Cambridge, MA*, 2003.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML-2001)*, 2001.

[5] K. Murphy. Dynamic Bayesian networks: Representation, inference and learning. *Ph.D. dissertation, UC Berkeley*, 2002.

[6] F. Peng and A. McCallum. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference (HLT-NAACL04)*, 2004.

[7] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[8] K. Sato and Y. Sakakibara. RNA secondary structural alignment with conditional random fields. In *Bioinformatics*, page 237C242, 2005.

[9] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *In Proceedings of HLT-NAACL*, page 213C220, 2003.

[10] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered HMMs. In *IEEE Transactions on Multimedia*, June 2006.