MULTI-MODAL LEARNING FOR GESTURE RECOGNITION

Congqi Cao, Yifan Zhang and Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100190
{congqi.cao, yfzhang, luhq}@nlpr.ia.ac.cn

ABSTRACT

With the development of sensing equipments, data from different modalities is available for gesture recognition. In this paper, we propose a novel multi-modal learning framework. A coupled hidden Markov model (CHMM) is employed to discover the correlation and complementary information across different modalities. In this framework, we use two configurations: one is multi-modal learning and multi-modal testing, where all the modalities used during learning are still available during testing; the other is multi-modal learning and single-modal testing, where only one modality is available during testing. Experiments on two real-world gesture recognition data sets have demonstrated the effectiveness of our multi-modal learning framework. Improvements on both of the multi-modal and single-modal testing have been observed.

Index Terms— multi-modality, gesture recognition, coupled hidden Markov model

1. INTRODUCTION

Traditional gesture recognition pertains to recognizing meaningful expressions of motion by a human, which is of utmost importance in designing an intelligent and efficient human-computer interface [1]. With the development of sensing devices, there are more and more multi-modal data (e.g. color image, 3-D depth image, audio, etc.) for gesture recognition in the real world. Therefore, how to utilize the multi-modal data effectively becomes one important problem. One common approach is to adopt multi-modal learning.

Multi-modal learning is becoming an increasingly essential task, since it can handle the issue of multi-modal information fusion and improve the performance of recognition [2]. Generally, there are three kinds of approaches to fuse multiple modalities together: early fusion (feature fusion), late fusion (decision fusion) and model-level fusion. Ngiam *et al.* proposed an application of deep networks to learn features over multiple modalities in [3]. Wu *et al.* used feature fusion approach in [4]. [3, 4, 5] all focused on feature learning. However, there is no explicit objective for these deep models to discover correlations across multiple modalities. And the results of feature fusion usually turn out to be most slightly inferior to decision fusion [6]. Wu *et al.* used decision fusion approach in [7] and got the best performance in the Multi-modal Gesture Recognition Challenge in 2013. However, decision fusion can not well discover the mid-level spatio-temporal interaction between different modalities, especially in long-term dependency sequential data. Thus, some research efforts resort to the model-level strategy to capture the mid-level correlation across multiple modalities.

Coupled hidden Markov model (CHMM) is an effective model to capture the dependency between multiple parallel threads. It has been used in many domains, such as audio-visual speech recognition [8], audio-visual emotion recognition [9], EEG classification [10], freeway traffic modeling [11] and action recognition [12, 13]. In the gesture recognition domain, most of the existing methods focus on using CHMM to model the interactions between two hands [12] or two arms [13]. To our best of knowledge, no one has attempt to use it to model the correlation between different modality data. Hence, we investigate the problem that if it is more effective to fuse multi-modal information in the CHMM than the strategy of feature-level and decision-level fusion. It is worthy noting that different from the work on audio-visual speech recognition [8], which also employ CHMM to couple the audio and visual signals, audio-visual gesture recognition is more challenging as the synchronization between the two modalities is not as well as the speech.

In this work, we propose a novel multi-modal learning framework, in which different modalities are coupled together in a unified graphical model. Currently, we employ a two-chain CHMM to implement the joint learning of two modalities. It is straightforward to extend the model to more modalities by coupling more HMM chains. In traditional multi-modal learning paradigm, multi-modal data is both available during the learning phase and the testing phase. Besides this, we employ a new paradigm: multi-modal data is utilized to learn a modality-shared model, while during testing, only one modality data is available to make the decision. This configuration does make sense in gesture recognition as during learning we can collect all modality data we have (e.g. audio and visual) to obtain a model; during testing, some modality data (e.g. audio) might not present in a real-world situation.

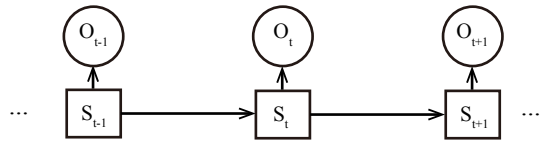

Fig. 1. First order HMM graphical model

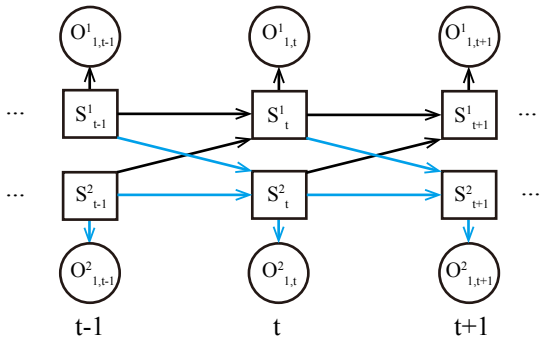

Fig. 2. Coupled HMM graphical model

Besides the traditional multi-modal learning paradigm, where the multi-modal information is both available during model training and testing, we also study a new learning paradigm, where the multi-modal information is available during model training, but only single-modal information is available during testing. The similar paradigm is also used by Vapnik et al. [14], which is called learning using privileged information (LUPI). The extra information used just during training stage functions as a "teacher" who can gave privileged knowledge on the training examples. However, our multi-modal learning and single-modal testing paradigm is still different to the "LUPI" problem. In "LUPI", privileged information is additional explanations about the raw data which is extracted from the same modality. For example, the privileged information explored in [15] for object classification is attribute, bounding boxes, image tags and annotator rationales. However, in our case, the data from different modalities individually carries one modality information for classification and recognition. Each modality data can train and test a classifier by itself.

2. FRAMEWORK

Hidden Markov model (HMM) shown as Figure 1 has been successfully used in perceptual computing for modeling and classifying dynamic behaviors. However, the standard form suffers from several limitations, such as not taking interactions among multiple chains into account. Therefore, it is not suitable to model correlations for multi-modal learning.

The coupled hidden Markov model (CHMM) [12] is a kind of dynamic Bayesian network that integrates two or more HMM chains where the discrete nodes at time t for each HMM are conditioned by the discrete nodes at time $t-1$ of all the related HMMs. It allows the hidden nodes of different HMM chains to interact, and to have their own observations. Figure 2 illustrates the CHMM used in our gesture recognition system. The squares represent discrete hidden states. The circles represent continuous observation nodes.

2.1. Parameters of CHMM

The parameters of CHMM are:

$$\lambda = (Q^C, O^C, A^C, B^C, \pi^C) \tag{1}$$

where C represents the number of coupled chains. Q^i and O^i are the possible hidden states and observations of channel i respectively. π^i represents the probability of initial states of channel i. A^C is the matrix of transition probabilities. B^C is the matrix of observation probabilities over states.

For two chains CHMM ($C = 2$), the parameters of the c-th chain at time t ($c \in \{1, 2\}$) can be represented as:

$$\pi^c(i) = p(q_1^c = i) \tag{2}$$
$$a^c_{i|j,k} = p(q_t^c = i | q_{t-1}^1 = j, q_{t-1}^2 = k) \tag{3}$$
$$b_i^c(o_t^c) = p(o_t^c | q_t^c = i) \tag{4}$$

Assuming the observation probabilities follow Gaussian distribution, $b_i^c(o_t^c)$ can be further written as:

$$b_i(o_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp[-\frac{1}{2}(o_t - \mu_i)^T \Sigma_i^{-1}(o_t - \mu_i)] \tag{5}$$

where μ_i and Σ_i are the mean and diagonal covariance of state i respectively. We omit the superscript c for simplification.

In the proposed CHMM, we assume that the observation probabilities of different chains are independent. The interaction between different modalities is modeled with the transition probabilities. Under the above assumption, observation probability at time t can be expressed by the product of component chains' observation probabilities.

$$b^{1,2}_{i,j}(o_t^{1,2}) = p(o_t^1 | q_t^1 = i)^\alpha p(o_t^2 | q_t^2 = j)^\beta \tag{6}$$

where α and β are the exponent weights in CHMM which adjust the importance between component chains.

2.2. Learning of CHMM

We extend Baum-Welch algorithm [16] in standard HMM for inference in CHMM. We recursively compute the forward $\alpha_t(i, j)$ and backward $\beta_t(i, j)$ variables for each combined state (i, j) and time t as follows:

$$\alpha_t(i,j) \triangleq p(q_t^1 = i, q_t^2 = j | o_{1:t}^{1,2}) \quad (7)$$

$$\boldsymbol{\alpha}_t \propto \boldsymbol{\varphi}_t \odot (\boldsymbol{\psi}^T \boldsymbol{\alpha}_{t-1}) \quad (8)$$

$$\beta_t(i,j) \triangleq p(o_{t+1:T}^{1,2} | q_t^1 = i, q_t^2 = j) \quad (9)$$

$$\boldsymbol{\beta}_{t-1} = \boldsymbol{\psi}(\boldsymbol{\varphi}_t \odot \boldsymbol{\beta}_t) \quad (10)$$

$$\varphi_t(i,j) = p(o_t^{1,2} | q_t^1 = i, q_t^2 = j) \quad (11)$$

$$\psi(i,j,k,l) = p(q_t^1 = i, q_t^2 = j | q_{t-1}^1 = k, q_{t-1}^2 = l) \quad (12)$$

where φ_t is the local evidence at time t, $\psi(i,j,k,l)$ is the transition matrix, the label $\odot$ represents the Hadamard product, T is the length of instance sequence, $\beta_T(i_{ter}, j_{ter}) = 1$ for terminal states i_{ter} and j_{ter}.

When estimating the parameters of CHMM using EM algorithm, we need to compute the expected statistics which can be obtained by running the forwards-backwards algorithms on each sequence. E step:

$$E[N_{k(1)}^1] = \sum_{i=1}^{N} p(q_{i1}^1 = k | o_i^1, \lambda^{old}) \quad (13)$$

$$E[N_j^1] = \sum_{i=1}^{N} \sum_{t=1}^{T_i} p(q_{i,t}^1 = j | o_i^1, \lambda^{old}) \quad (14)$$

$$E[N_{ijk}^1] = \sum_{l=1}^{M_2} \sum_{i=1}^{N} \sum_{t=2}^{T_i} p(q_{i,t-1}^1 = i, q_{i,t-1}^2 = j, q_{i,t}^1 = k, q_{i,t}^2 = l | o_i^{1,2}, \lambda^{old}) \quad (15)$$

$$\gamma_t(i,j) \triangleq p(q_t^1 = i, q_t^2 = j | o_{1:t}^{1,2}) \quad (16)$$

$$\gamma_t(i,j) \propto \alpha_t(i,j)\beta_t(i,j) \quad (17)$$

$$\xi_{t,t+1}(i,j,k,l) \triangleq p(q_t^1 = i, q_t^2 = j, q_{t+1}^1 = k, q_{t+1}^2 = l | o_{1:t}^{1,2}) \quad (18)$$

$$\xi_{t,t+1}(i,j,k,l) \propto \psi \odot (\alpha_t (\phi_{t+1}\beta_{t+1})^T) \quad (19)$$

In the above equations, $E[N_{k(1)}^1]$ stands for the expected counts for state k as the first hidden node of chain 1. $E[N_j^1]$ is the expected counts for state j of chain 1. $E[N_{ijk}^1]$ is the expected counts for transiting to state k of chain 1 from combined state (i,j). N is the number of instances. T_i is the length of instance i. M_2 is the number of hidden states of chain 2 . ϕ represents the observation probability.

The estimated parameters of chain 1 can be computed with expectations obtained in E step. Chain 2 follows the same way. For simplification, we omit the superscript in equations. M step:

$$\hat{A}_{ijk} = \frac{E[N_{ijk}]}{\sum_{k'} E[N_{ijk'}]} \quad (20)$$

$$\hat{\pi}_k = \frac{E[N_{k(1)}]}{N} \quad (21)$$

$$\hat{\mu}_k = \frac{E[\bar{o}_k]}{E[N_k]} \quad (22)$$

$$\hat{\Sigma}_k = \frac{E[\bar{o}_k \bar{o}_k^T] - E[N_k]\hat{\mu}_k \hat{\mu}_k^T}{E[N_k]} \quad (23)$$

where $E[\bar{o}_k]$ and $E[\bar{o}_k \bar{o}_k^T]$ can be computed with $\gamma_t(i,j)$ similar to [16].

For standard HMM, Baum-Welch algorithm requires calculation of a forward and a backward variables for each state i and time t. Thus it needs to calculate $N * T$ variables and each one requires $O(N)$ time giving an overall complexity of $O(TN^2)$. We use a direct generalization of this algorithm to CHMM, which requires calculation of forward and backward variables of the coupled channels $1 \cdots C$ as expressed in Equation (7) to (10). Thus we need to compute $N^C * T^C$ variables and hence the procedure is inherently exponential in C. However there are many researches aiming to decrease the computational complexity of CHMM [11, 17] and CHMM can be easily paralleled to speed up. We leave this to the future work and stick with naive implementation in this paper.

2.3. Recognition in CHMM

By combining all the individual CHMMs trained for every gesture class together, we get a multi-class classifier. We classify testing instance by computing the observation probability of the whole sequence with every CHMM trained for each gesture class. Then we choose the most likely class label. We use l stands for the class label index.

$$l^* = \arg \max_{1 \le l \le L} p(o | \lambda^l) \quad (24)$$

$$p(o|\lambda^l) = \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} \alpha_T^l(i,j) \quad (25)$$

The time complexity of single-modal testing is the same as standard HMM. For single-modal testing, we need to solve the problem of missing observations of the other modality. We transfer the combined transition probabilities to single chain transition probabilities as shown in Figure 3 following the equations as below:

$$p(q_{t+1}^1 | q_t^1) = \frac{p(q_t^1, q_{t+1}^1)}{p(q_t^1)} \quad (26)$$

$$= \frac{\sum_{q_t^2} p(q_t^1, q_t^2, q_{t+1}^1)}{p(q_t^1)} \quad (27)$$

$$= \frac{\sum_{q_t^2} p(q_{t+1}^1 | q_t^1, q_t^2) p(q_t^1, q_t^2)}{p(q_t^1)} \quad (28)$$

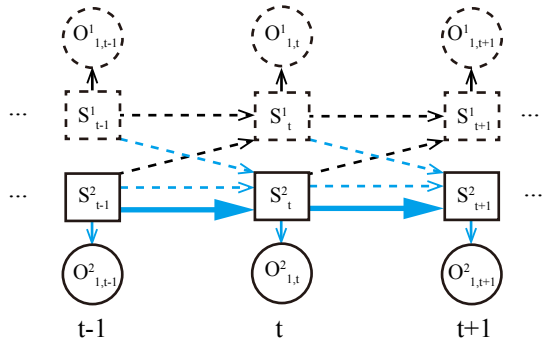

Fig. 3. The illustration of how to transform CHMM to HMM

Table 1. Notations used in the experimental result tables

Notations	Descriptions
HMM feature fusion	Learning and testing HMM with concatenated features of different modalities
HMM decision fusion	Fusing the decision scores from each HMM trained by different modalities separately
CHMM	Learning and testing CHMM with multi-modal data
HMM (*)	Learning and testing HMM with features of (*) modality data
CHMM (*) test	Learning CHMM with multi-modal data Tesing with (*) modality data

3. EXPERIMENTS

We evaluate our methods on two public gesture recognition data sets with three kinds of modality combinations. On ChaLearn MMGR data set, we use depth-color combination and audio-skeleton combination. On ChaAirGest data set, we use xsen-skeleton combination. The beginning and ending points of each gesture interval in the videos are all provided. Thus the gesture interval segmentation does not need to be performed. The gesture recognition can be just considered as a classification problem. However, with the aim to compare our method with the two state-of-the-art works [4, 7], in which interval segmentation and classification are both performed, we also automatically determine the beginning and ending points of each gesture interval without using the manual labels. In our CHMM, the number of hidden states for each modality is set to be 10, and CHMM is trained with 20 iterations in all experiments. We list the notations of compared methods in Table 1.

3.1. Experiment 1: ChaLearn MMGR 2014

A challenge on multi-modal gesture recognition was organized in 2013 and 2014 [18]. A large video database of 20 Italian gesture categories (such as "perfetto" or "ok") is available . The skeletal model, user mask, RGB and depth images captured by the Kinect sensor are provided. We use the Development and Validation data sets which contain groundtruth labels for training and testing respectively.

Table 2. Classification accuracy on ChaLearn MMGR 2014 data set with depth and color data

Method	Overall	Mean	Std	T_{test}
HMM feature fusion	0.7032	0.7010	0.1222	180.2331
HMM decision fusion	0.7264	0.7242	0.1345	178.4876
CHMM	**0.7351**	**0.7328**	0.1300	814.2878
HMM depth	0.6763	0.6735	0.1527	111.7195
CHMM depth test	**0.7001**	**0.6979**	0.1327	94.7953
HMM color	0.5906	0.5890	0.1285	88.8202
CHMM color test	**0.6031**	**0.6007**	0.1229	99.9768

We use color modality and depth modality of ChaLearn MMGR 2014 data set to carry out experiments. There are 6850 instances for training and 3454 instances for testing. We first extract HOG features from color and depth video respectively, then use the features to compute a dictionary in order to obtain bag of word features. After PCA process, features of 350-dimension are used as the input observations of CHMM.

We compare our CHMM method for multi-modal testing with feature fusion method and decision fusion method which are introduced as baselines in [19]. Furthermore, we compare our method for single-modal testing with HMMs. Since the numbers of testing instances that belong to each gesture class are different, we compute the overall accuracy and mean accuracy of 20 classes respectively. The overall accuracy is the weighted summation of the accuracy values of all 20 classes, where the weights are proportional to the size of the classes. The mean accuracy is the numerical average of the accuracy values of all 20 classes. We also list the time cost for testing in seconds running on Intel i7-3770 CPU @3.4GHz.

As shown in Table 2, when we use the traditional multi-modal learning paradigm, where the multiple modalities are both available during CHMM training and testing, our model achieves a higher classification accuracy than the feature-fusion and decision-fusion methods. It demonstrates that coupling multiple modalities in model-level is more effective to capture the intrinsic dependencies between modalities and thus better to model the gesture. When we use the new learning paradigm, where only one modality data is available during testing, it can be seen that the performance is also superior than the HMM which trained by a single modality (i.e. the depth feature or the color feature). This verifies that in training phase, the extra modality data can help to better learn the model parameters. Even though it is missing during testing,

Table 3. Classification accuracy on ChaLearn MMGR 2013 data set with audio and skeleton data

Method	Overall	Mean	Std	T_{test}
HMM feature fusion	0.7149	0.7104	0.1803	34.3706
HMM decision fusion	**0.9473**	**0.9466**	0.0361	43.5545
CHMM	0.9405	0.9396	0.0454	911.3699
HMM audio	0.8951	0.8953	0.0680	24.0758
CHMM audio test	**0.8957**	**0.8962**	0.0440	46.5079
HMM skeleton	0.6149	0.6089	0.2219	24.0814
CHMM skeleton test	**0.6305**	**0.6236**	0.2178	25.5096

Table 4. Classification accuracy on ChaAirGest data set with Xsens and skeleton data

Method	Mean	Std	T_{test}
HMM feature fusion	0.9085	0.0506	9.0931
HMM decision fusion	**0.9184**	**0.0576**	3.8940
CHMM	0.9109	0.0494	86.8990
HMM xsens	0.8974	0.0627	2.6649
CHMM xsens test	**0.9042**	**0.0441**	3.1800
HMM skeleton	0.6371	0.1144	1.6916
CHMM skeleton test	**0.6866**	**0.0945**	2.4154

it still enhances the performance of the model.

3.2. Experiment 2: ChaLearn MMGR 2013

In ChaLearn MMGR 2013 data set, the audio modality data is provided. In this experiment, we employ the audio-skeleton modality combination. There are 7205 instances for training and 3280 instances for testing since we filter out gesture instances which contain invalid all-zero skeleton data.

We choose Mel frequency cepstral coefficients (MFCCs) as our audio features. 25 ms Hamming window with 10 ms shift is used to compute 39-dimension MFCC features. We perform end-point detection in order to remove non-speech intervals just like [7]. Firstly we compute the average short-time energy. Then we set thresholds to determine the beginning and end of a gesture instance. As for skeleton features, we only use the original 3D coordinates of human bodys joint points provided by the data set to obtain the relative 3D position of upper body joint points which defined on directly connected joint pairs. The skeleton features contain 36 dimensions. As the sampling rates are different between the audio data and the skeleton data, in order to align the two modalities, we down sample the audio data to make the length of the feature sequence the same as the skeleton feature sequence.

Although in Table 3, the traditional multi-modal learning paradigm, where the multiple modalities are both available during CHMM training and testing, is slightly inferior to HMM decision-fusion method, there is a significant improvement compared with HMM feature-fusion method. And our new learning paradigm, where only one modality data is available during testing is effective. The performance is superior than the HMM which trained by a single modality (i.e. the audio feature or the skeleton feature). This verifies that in the training phase, the extra modality data can help to better learn the model parameters. The performance is mainly restricted by the asynchronous of audio and skeleton data. If we could utilize a more appropriate alignment method, there will be an improved result of CHMM.

3.3. Experiment 3: ChaAirGest

The corpus containing 10 different gestures with 1200 gesture instances is provided by a challenge for multi-modal mid-air gesture recognition in 2013 [20]. This data set is captured with a Kinect camera and four body-worn inertial motion units (IMU). Each Xsens IMU sensor can provide linear and angular acceleration, magnetometer, Euler orientation, orientation quaternion and barometer data with 50 Hz frequency.

We use the data captured by Xsens IMU sensor and the skeleton data captured by Kinect to conduct the experiment. Since there is no individual training and testing data sets, we perform leave-one-out cross validation. In each round, one-tenth gesture instances are used for testing (120), others are for training (1080). Totally there are ten rounds. Because the numbers of testing instances of 20 gesture classes are all the same, overall accuracy is the same as mean accuracy of 20 gesture classes. We only list mean accuracy in Table 4.

Raw data collected by two of the four Xsens is used as Xsens features which is 34-dimension. Skeleton features are extracted as Experiment 2. Because the positions of hip-center and spine are not tracked in this data set, the skeleton features are 33-dimension. We align Xsens data to be the same length as skeleton by sampling.

The performance of CHMM is also restricted by the asynchronous of data. Results in Table 4 demonstrate that our new learning paradigm, where only one modality data is available during testing, does improve the performance of single modality recognition compared with HMM trained by a single modality (i.e. the xsens feature or the skeleton feature).

3.4. Experiment 4: ChaLearn MMGR 2013

In order to compare to the two state-of-the-art works [4, 7] which perform both of the gesture interval segmentation and

Table 5. Classification accuracy on ChaLearn MMGR 2013 data set with audio and skeleton data after automatic gesture interval segmentation

Method	Classification accuracy
Multi-modal DBN +HMM[4]	0.701
HMM decision fusion	0.7979
CHMM	**0.8029**
skeleton DTW[7]	0.4434
skeleton HMM	0.5079
CHMM skeleton test	**0.5409**

gesture classification, we also automatically detect the beginning and ending points of each interval by using the start and end time of each audio fragment to extract the corresponding skeleton fragment. In [4], dynamic Deep Belief Networks are deployed to extract the feature representations of high level audio and skeletal joints in a feature fusion framework. Since Wu *et al.* only uses the Development data set for training and testing, we also use 393 labeled sequences with totally 7754 gestures, in which 350 sequences for training and 43 sequences for testing to keep the experimental condition same. F1 score is used as the measurement.

As shown in Table 5, compared with feature fusion method in [4] and skeletal joints based method in [7] using Dynamic Time Warping method, our multi-modal learning paradigms show significant improvements.

4. CONCLUSION

This paper presents a novel coupled hidden Markov model framework to explicitly model the interaction between chains for gesture recognition with two configurations: one is multi-modal learning and multi-modal testing, the other is multi-modal learning and single-modal testing. Instead of feature fusion or decision fusion strategies which are difficult in exploiting the correlation between multiple modalities, we resort to model-level fusion strategy. The experimental results show that our approach is appropriate for multi-modal learning problem, especially when only one modality is present at test time which is meaningful for practical applications.

5. ACKNOWLEDGEMENTS

We thank the anonymous reviews for their valuable comments. This work was partly supported by 863 Program (2014AA015104) and National Natural Science Foundation of China (61332016, 61202325, 61379100).

6. REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *T SYST MAN CY C*, vol. 37, no. 3, pp. 311–324, 2007.

[2] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools.," in *ICME*. 2013, pp. 1–6, IEEE.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and Andrew Y Ng, "Multimodal deep learning," in *ICML*, 2011, pp. 689–696.

[4] D. Wu and L. Shao, "Multimodal dynamic networks for gesture recognition," in *MM*. 2014, pp. 945–948, ACM.

[5] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *JMLR*, vol. 15, pp. 2949–2980, 2014.

[6] Cees G. M. Snoek, "Early versus late fusion in semantic video analysis," in *MM*. 2005, pp. 399–402, ACM.

[7] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," in *ICMI*. 2013, pp. 453–460, ACM.

[8] A.V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and Kevin Murphy, "A coupled hmm for audio-visual speech recognition," in *ICASSP*, May 2002, vol. 2, pp. II–2013–II–2016.

[9] K. Lu and Y. Jia, "Audio-visual emotion recognition with boosted coupled hmm," in *ICPR*, Nov 2012, pp. 1148–1151.

[10] S. Zhong and J. Ghosh, "Hmms and coupled hmms for multi-channel eeg classification," in *IJCNN*, 2002, vol. 2, pp. 1154–1159.

[11] J. Kwon and K. Murphy, "Modeling freeway traffic with coupled hmms," Tech. Rep., University of California, Berkeley, 2000.

[12] M. Brand, N. Oliver, and A. Pentland, "Coupled hidden markov models for complex action recognition," in *CVPR*, Jun 1997, pp. 994–999.

[13] H. Ren and G. Xu, "Human action recognition with primitive-based coupled-hmm," in *ICPR*, 2002, vol. 2, pp. 494–498 vol.2.

[14] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009.

[15] V. Sharmanska, N. Quadrianto, and C.H. Lampert, "Learning to rank using privileged information," in *ICCV*, Dec 2013, pp. 825–832.

[16] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[17] P. Natarajan and R. Nevatia, "Coupled hidden semi markov models for activity recognition," in *WMVCW*, Feb 2007, pp. 10–10.

[18] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, Is. Guyon, V. Athitsos, and H. Escalante, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *ICMI*. 2013, pp. 445–452, ACM.

[19] J. Wu and J. Cheng, "Bayesian co-boosting for multi-modal gesture recognition," *JMLR*, vol. 15, pp. 3013–3036, 2014.

[20] S. Ruffieux, D. Lalanne, and E. Mugellini, "Chairgest: A challenge for multimodal mid-air gesture recognition for close hci," in *ICMI*. 2013, pp. 483–488, ACM.