基于语音视觉特征身份鉴别的层级结构

吴志勇　蔡莲红　王志明
清华大学计算机系，北京，100084
wuzy99@mails.tsinghua.edu.cn

摘　要：　本文针对语音和视觉特征的动态关联特性，将生物特征按照层级模型来加以描述，并基于此提出了一种身份鉴别的层级结构，将身份鉴别按照其融合的过程分成不同的级别：数据级、参数级、特征级、模型级、和决策级。该层级结构基于语音和视觉特征提出，但是不一定非得局限于此，而是可以拓展到任意的生物特征的鉴别过程中。

关键词：　身份鉴别，数据融合，层级结构，模态

1. 引 言

生物身份鉴别包括语音、脸像、指纹、虹膜、掌纹、签字、唇动（唇部运动）等多种特征技术。其中，指纹、虹膜、掌纹等特征被认为终生不变，对其研究较多，技术也相对比较成熟。而语音、脸像、唇动等富于变化，对其识别和描述相对比较困难，是目前研究的焦点。

在语音和视觉的研究中，现有的研究多采用单模态的方法，如脸像识别、说话人识别等。实际上，人们对语言的理解是多模态的，耳朵听辨声音的同时，眼睛会去观察说话人的面部表情，而说话时复杂多变的面部表情不仅可以传达丰富的感情，而且可以增强对语言的理解。因此，近年来，数据融合技术，即通过一定的综合策略将多种模态结合起来进行身份鉴别开始得到重视。其中一个著名的系统是 BioID, 它使用了脸像、语音和唇动三个特征来识别身份[1]。

本文关注的是多模态身份鉴别中数据融合及其层级结构方面的问题。目前，多模态的身份鉴别在数据融合方面基本上都使用了一种简化的策略：各个模态分别独立进行处理，比如首先进行特征的提取、模型的建立和匹配，然后将单模态匹配的结果通过一定的方法在决策阶段进行融合，并得到最终的综合判决结果。这种策略处理简单，但是没有充分考虑不同特征之间的关联关系，而这种关系在鉴别过程的不同层次阶段具有不同的内在内容。通过在不同层级阶段对关联关系进行研究，将会对鉴别过程有更深入的了解，并提高鉴别的性能。

国外的研究机构对数据融合的层级方面进行了一定的研究。MSU 的研究者们将数据融合分为三个层级：（1）特征抽取层（Feature extraction level），不同模态抽取的特征进行拼接融合，然后进行模板的匹配和决策；（2）决策层（Confidence level），不同模态分别进行匹配，得到的中间结果进行融合；（3）抽象层（Abstract level），对单个模态分别进行决策，将决策结果通过融合模块进行接受或者拒绝的二元决策融合[2]。2002 年第一期的 IEEE 多媒体杂志（IEEE transaction on Multimedia）上关于以语音为基础的双模态识别综述中给出了特征融合和决策融合的系统结构，对于特征融合和决策融合作了简单的说明[3]。

本文针对生物特征的特点，将生物特征按照层级模型来加以描述，并基于此提出了一种身份鉴别的层级结构，将身份鉴别按照其融合的过程分成不同的级别：数据级、参数级、特征级、模型级、和决策级。文中主要说明了生物特征的分层描述模型以及身份鉴别的层级结构；然后结合语音视觉特征，集中于参数级的融合阶段，介绍了我们的初步研究工作，详细说明了语音视觉之间的动态关联关系；最后给出了总结和讨论。

2. 身份鉴别的层级结构

2.1 生物特征分层描述模型

不同的生物特征具有各自不同的特点，也具有不同的内在内容，但是从整体上看，不同

的生物特征，包括其数据的表示以及研究的方法等，都可以用同样的层级模型来加以描述。生物特征的分层描述模型包括四个层级：最低层的数据级、次低层的参数级、次高层的特征级、最高层的模型级。从低级到高级，是一个数据逐级抽象、数据表示逐级概括的过程。

图 1 示意说明了生物特征的层级描述模型，右侧以语音特征和视觉特征为例，说明了每个层级的可能包括的内容。

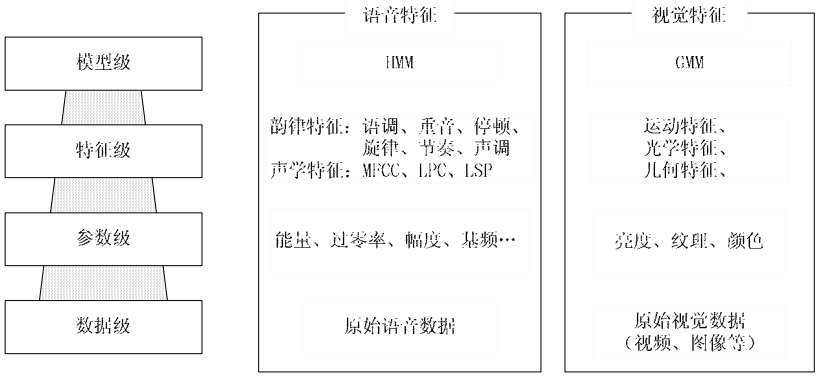

图 1　　生物特征分层描述模型

以语音特征为例，在最低层的数据级，直接包括了对原始语音数据的一种表示，比如录音数据等，它携带了对于语音特征的最原始、最充分的描述。

下一个较高的层级是参数级，该层级所处理的数据是原始语音数据经过一定的处理以后得到的一些语音参数，比如描述语音音量大小的幅度、能量等，描述波动特性的过零率、基频等。参数级的数据对原始数据进行了抽象概括。

在更高层的特征级，既可以有描述语音声学特征的参数，比如 MFCC、LPC、LSP 参数等等，也包括了用于描述对语音的感知的韵律特征参数，比如语调、重音、停顿等。

最高层是模型级，对数据的产生、描述和表示进行模型的建立，在语音中用得最普遍的是 HMM 模型。在生物特征鉴别中，可以对模型进行匹配比较，产生鉴别结果。

从生物特征的分层描述模型可以看出：越低层，参数越具体、处理越复杂；越高层，参数越抽象、处理越简单。对于特征鉴别和数据融合来说，较低的层级，考虑参数之间的关联特性将会使得信息更加充分完备，但是数据相关性高，处理复杂；较高的层级，数据相关性降低，处理简单，但是忽略了参数之间的联系，造成了信息的丢失。

2.2 数据融合的层级结构

生物特征鉴别对生物特征进行处理，利用匹配、分类的方法，将待鉴别的特征数据进行归类。多生物特征鉴别技术，将不同单个特征处理的结果加以综合，考虑不同特征从多个方面得到的信息，从而使得鉴别得到更好的结果。这一过程可以通过数据融合技术加以实现。

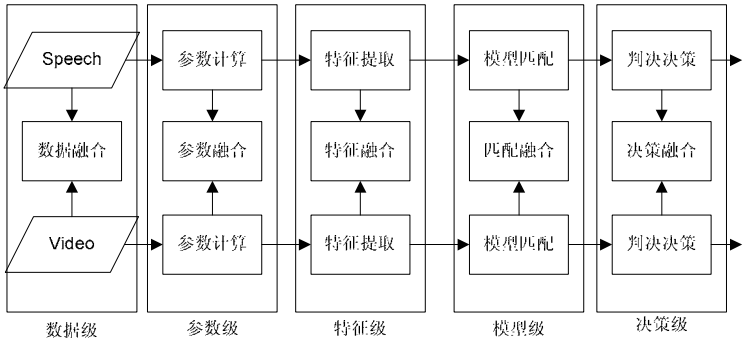

图2　　多层级融合框架结构

相应于生物特征的分层描述模型，数据的融合处理可以在模型的不同层次级别上进行，因此数据的融合也可以有不同的层级，图 2 描述了数据融合的层级框架结构。

2.2.1 数据级融合

最低层的数据级，直接包括了对原始特征最充分、最有效的描述，然而对于计算机处理而言，由于数据的大量性、特征的复杂性、以及数据之间的强相关性等，使得直接利用原始数据的融合几乎是不可能。因此数据级的融合还只能是层级框架的一个理论组成部分，距离实际应用还有较大的距离。

2.2.2 参数级融合

特征的提取是进行生物鉴别的基础和关键，而参数的计算又是特征提取的根本。不同的特征采用不同的参数计算方法。然而某些特征之间存在关联关系，比如说话人识别时语音和脸像、唇动之间。现有的主要研究成果，都是分别进行语音、脸像、唇动的参数计算和特征提取，而实际上通过语音与脸像、唇动之间关联关系的研究，考察他们之间的关联关系，可以对参数的提取提供辅助作用。而另一方面，通过关联关系的研究并建立联合模型，并基于此模型作为匹配和分类器的输入，从而可以直接用来进行身份的鉴别。

2.2.3 特征级融合

输入数据经过前端处理以后，对于每种生物特征分别得到其特征描述向量，然后经过特征融合的处理，将多个低维的特征描述向量融合（合并）形成更高维的联合特征向量参数。在系统建模时，针对该高维向量进行模型的建立，匹配时，利用该高维向量进行匹配。

2.2.4 模型级融合

对特征进行建模时，可以对单个的特征分别建立模型，也可以同时考虑多个特征建立联合模型，而后端的分类器基于此联合模型进行身份的鉴别。

联合模型的建立，既可以考虑多个特征之间的关联，也可以考虑他们之间的区别特性。比如对于语音和视觉特征的处理，从关联关系来说，可以考虑建立耦合 HMM（Coupled HMM）模型，如图 3 所示；而另一方面，可以通过建立混淆树模型，来描述他们之间的区别特性。

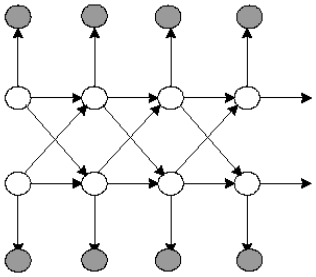

图 3　　耦合 HMM（Coupled HMM）模型

2.2.5 决策级融合

决策融合的系统，输入信号经过处理得到特征参数，然后分别进行单模态的建模与识别，将识别的中间结果参数，通过决策融合模块进行融合，然后通过多模态决策算法得到最终的鉴别结果。

2.3 数据融合的研究现状

数据融合通过一定的综合策略将不同单个特征模块处理得到的结果加以融合以综合考虑不同的信息从而得到更好的结果。

目前关于数据融合的研究主要集中在决策融合阶段[1][2]。其特点首先是简单可行，不

同的单个特征可以分别进行独立的处理，比如参数的计算、特征的提取、模型的建立等，然后进行模型的匹配，得到匹配的分值，最后通过决策融合的过程，将多个匹配结果经过一定的融合算法进行综合，得到最终的结果。而融合算法的处理，可以转化为模式识别的过程，基本上很少涉及到特征本身。而另一方面，在决策阶段进行融合，可以方便的进行特征的扩充或者删减，只需要增加或减少融合匹配的输入参数的个数即可，而不会对整个系统的结构产生影响。

但是决策融合阶段的研究，其缺点也是显而易见的，虽然使用了多个特征的综合结果以提高系统的性能，但是在处理过程中，在较低层级上各个特征之间完全独立，忽视了特征之间的关联关系所带来的作用和影响，另外，决策融合主要集中于融合算法的讨论，而忽视了对于生物特征的更多考虑。

因此需要在较低层级上，比如参数级、特征级等层面上研究数据的融合策略。

3． 语音视频特征的参数级融合模型

语音是最为自然的人机交互方式，用来进行身份的鉴别较易被用户所接受。语音的发声与嘴唇的运动以及脸像特征的变化都存在必然的联系。因此，有必要研究嘴唇运动伴随语音变化的关系，以及脸像特征信息（如头部运动、面部光流场、几何以及纹理变形等）所提供的丰富的多层次的信息，并利用这些信息辅助语音进行身份的鉴别。

3.1 唇动特征与语音变化关系

我们初步研究了视频脸像，特别是嘴部运动，和语音发声之间的关系。图4说明了利用视位（Viseme）信息的唇动特征提取和建模方法。通过研究发现，由于语音发声的时序关系，唇部运动特征最为强烈明显的部分（口形的形成）在声音发出之前已经形成，然后逐渐根据发音的情况而发生变化，图5说明了口形的形成和语音变化之间的时序关系。

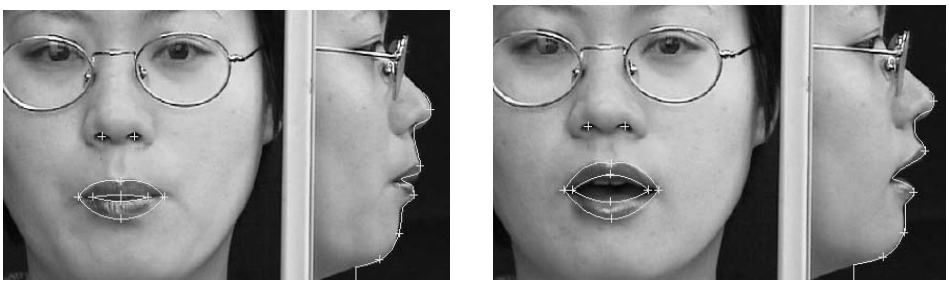

图 4　　唇动特征的提取和建模

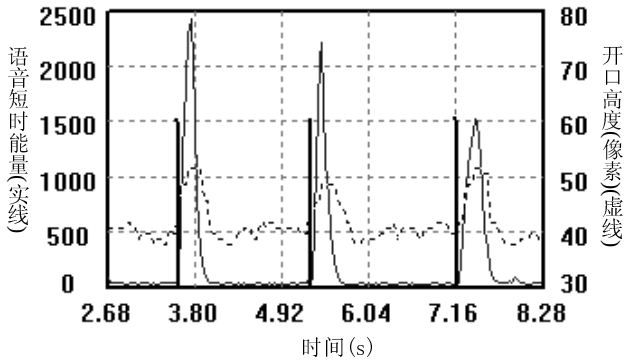

图 5　　口形与语音变化的时序关系

3.2 语音－视频联合模型

为了进一步研究语音、视频特征之间的关系，我们提出了语音－视频联合模型，如图6

所示。通过该模型的建立，使得我们对于如何研究语音与脸像、唇动特征之间的关系有了更加深刻的认识，指明了研究的方法。从图中可以看出，我们可以从如下几个方面来进行研究：

（1）语音端点检测与镜头切分之间的关联，建立语音与图像的时序同步模型；

（2）研究语音切分与唇动的映射关系，建立基于唇形和唇动信息的语音分帧算法；

（3）研究说话人头像运动规律和语音韵律边界间的关系，改善头像跟踪定位检测算法；

（4）研究韵律参数凸显（重音、情感等）对唇动、脸像的影响，建立语音与唇动脸像的语义关联模型。

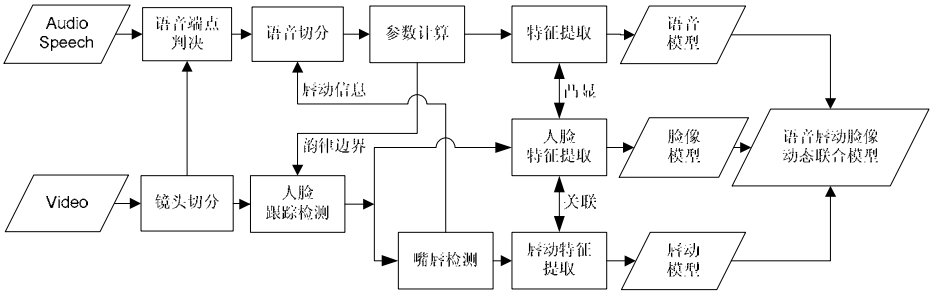

图 6　　语音、视频联合模型关系描述

4. 总结与讨论

本文针对语音和视觉特征的动态关联特性，提出了生物特征的层级模型，并基于此提出了一种身份鉴别的层级结构，将身份鉴别按照其融合的过程分成不同的级别：数据级、参数级、特征级、模型级、和决策级。该层级结构基于语音和视觉特征提出，可以拓展到任意的生物特征的鉴别过程。

通过对融合过程的层级划分，为研究工作提供一种结构框架，使得每个部分相对独立，拥有各自的模块，从而可以集中精力研究某个层级内部的表现，探讨研究内容，试验研究方法和策略，另外通过明确每个层级的不同的特点，可以针对不同的层级采用不同的方法，研究不同的内容。

语音视频特征的参数级融合模型，详细说明了语音、视频之间的关联关系，是今后开展进一步研究的方向。

参考文献：

[1] Frischholz, R.W. and U. Dieckmann, BioID: A Multimodal Biometric Identification System. Computer, 2000. 33(2): p. 64-68

[2] Ross, A., A. Jain, and J.-Z. Qian. Information Fusion in Biometrics. in Proc. of 3rd Int'l Conference on Audio- and Video-Based Person Authentication. 2001. Sweden.

[3] Chibelushi, C.C., F. Deravi, and J.S.D. Mason, A Review of Speech-Based Bimodal Recognition. IEEE transaction on multimedia, 2002. 4(1): p. 23-37.

[4] 王志明，蔡莲红. 汉语语音视位的研究. 应用声学. 2002年5月. 第21卷第3期. p.29-34