



[12] 发明专利申请公开说明书

[21] 申请号 01807753.6

[43] 公开日 2003 年 6 月 4 日

[11] 公开号 CN 1422494A

[22] 申请日 2001.12.3 [21] 申请号 01807753.6

[30] 优先权

[32] 2000.12.5 [33] US [31] 09/730,204

[86] 国际申请 PCT/EP01/14275 2001.12.3

[87] 国际公布 WO02/47386 英 2002.6.13

[85] 进入国家阶段日期 2002.10.8

[71] 申请人 皇家飞利浦电子有限公司

地址 荷兰艾恩德霍芬

[72] 发明人 S·古塔 H·J·斯特鲁贝

A·科梅纳雷兹

[74] 专利代理机构 中国专利代理(香港)有限公司

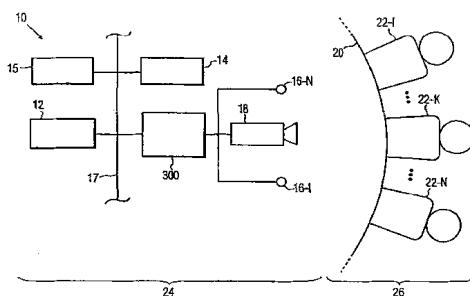
代理人 杨凯陈霖

权利要求书 3 页 说明书 16 页 附图 5 页

[54] 发明名称 在电视会议和其他应用中预测事件的方法和装置

[57] 摘要

本发明公开了利用声音和视觉线索预测事件的方法和装置。本发明处理音频和视频信息来识别一个或多个(1)声音线索、例如语调类型，音调与音量；(2)视觉线索、例如注视目光、面部姿势、身体姿势、手势以及面部表情；或(3)上述线索的组合；这些线索通常与一个事件相关联，例如电视会议的与会者在发言前表现的行为。这样本发明就能使视频处理系统预测事件，例如识别下一个发言人。预测发言人识别器以一种学习模式工作，根据在有或没有一个或多个预定义的声音或视觉线索的情况下与会者“会发言”或“不会发言”的概念来学习每个与会者的特征信息。预测发言人识别器以预测模式工作，将特征信息中嵌入的学习特征与音频和视频信息加以比较，从而预测下一个发言人。



1. 一种利用音频和视频信息中的至少一种信息预测事件的方法，所述方法包括以下步骤：

5 建立定义能提示某一既定事件的行为特征的多个线索；以及
处理所述音频和视频信息中的至少一种信息来识别一种所述线索 (410, 420)。

2. 如权利要求 1 所述的方法，其特征在于：所述多个线索包括标识一个人通常在发言之前表现出的行为的至少一种线索。

10 3. 如权利要求 1 所述的方法，其特征在于：所述多个线索包括标识一个人通常在结束发言之前表现出的行为的至少一种音频线索。

4. 如权利要求 1 所述的方法，其特征在于还包括获得与所述标识的线索有关联的所述某人的图像的步骤。

15 5. 如权利要求 1 所述的方法，其特征在于还包括保持至少一个人的简档 (500) 的步骤，所述简档 (500) 建立关于所述多个线索中一个或多个线索的阈值。

6. 一种在视频处理系统 (300)(10) 中跟踪发言人 (22-k) 的方法，所述视频处理系统 (300)(10) 处理音频和视频信息中至少一种信息，所述方法包括以下步骤：

20 处理所述音频和视频信息中至少一种信息、以便识别定义提示一个人即将发言的行为特征的多种线索中至少一种线索；以及
获得与所述识别的线索有关联的所述某人的图像。

7. 如权利要求 6 所述的方法，其特征在于：至少一台摄像机 (18) 按照与所述线索有关联的人所关联的全景、倾斜和变焦值来聚焦。

25 8. 如权利要求 6 所述的方法，其特征在于：所述多个线索包括标识一个人通常在发言之前表现出的行为的至少一种音频线索。

9. 一种利用音频和视频信息中的至少一种信息来预测事件的系统 (300)，它包括：

存储计算机可读代码的存储器；以及
工作时连接到所述存储器的处理器；所述处理器配置成：
建立定义能提示某一既定事件的行为特征的多个线索；以及
处理所述音频和视频信息中的至少一种信息来识别一种所述线
索。

10. 一种用于跟踪视频处理系统(300)(10)中的发言人(22-k)
的系统(300)，所述视频处理系统(300)(10)处理音频和视频信息
中至少一种信息，它包括：

存储计算机可读代码的存储器；以及
工作时连接到所述存储器的处理器；所述处理器配置成：
处理所述音频和视频信息中至少一种信息、以便识别定义提示一
个人即将发言的行为特征的多种线索中至少一种线索；以及
获得与所述识别线索有关联的所述某人的图像。

11. 一种利用音频和视频信息中至少一种信息来预测事件的制造
品，它包括：

在其上实现计算机可读代码工具(mesns)的计算机可读介质，所述
计算机可读程序编码工具包括：

建立定义能提示某一既定事件的行为特征的多个线索的步骤；以
及

20 处理所述音频和视频信息中至少一种信息、以便识别一种所述线
索的步骤。

12. 一种用于跟踪视频处理系统(300)(10)中的发言人(22-k)
的制造品，所述视频处理系统(300)(10)处理音频和视频信息中至
少一种信息，它包括：

25 在其上实现计算机可读代码工具(mesns)的计算机可读介质，所述
计算机可读程序编码工具包括：

处理所述音频和视频信息中至少一种信息、以便识别定义提示一
个人即将发言的行为特征的多种线索中至少一种线索的步骤；以及

获得与所述识别线索有关联的所述某人的图像的步骤。

在电视会议和其他应用中预测事件的方法和装置

5 发明领域

本发明一般涉及视频信号处理领域，更具体地说，涉及预测事件的技术，例如在视听表演中（例如电视会议）预测下一个发言人。

发明背景

10 电视会议的应用日渐增加，它使远距的用户不仅在声音上，而且在视觉上相互交流。这样，即使远距的用户不能身在同处，但电视会议系统使用户好象在同一房间内进行交流，让他们通过可见的姿势和面部表情来更好地表达自己的意思。通过最终输出的视频信号跟踪某一与会者，是电视会议系统的一个重要方面。

15 电视会议系统常利用摇俯变焦(pan-tilt-zoom) (PTZ) 摄像机来跟踪当前的发言人。通过 PTZ 摄像机，系统定位并进行光学变焦以完成跟踪任务。最初，电视会议系统中 PTZ 摄像机的控制系统要求操作者对摄像机作手动调节，才能始终聚焦在当前发言人的身上。但电视会议系统的用户对非手工操作的要求日益强烈，这使得 PTZ
20 摄像机的控制必须是全自动的。

人们已提出多种根据音频和视频信息来自动检测人物的方法。音频定位器通常用来处理从话筒阵列中获得的音频信息，以确定发言人的位置。具体地说，当知道各相对的话筒位置时，就可以用众所周知的三角技术来估计距单一源的声波传播时间差，从而测定声源的位置。同理，视频定位器通常定位视频图像中一个或多个关心的对象，例如在电视会议中发言人的头和肩膀。有许多已知的技术可用来检测图像中一个人的位置，例如，在“面部识别：从理论到应用”（NATO ASI Series, Springer Verlag, New York, H. Wechsler et al., editors, 1998）中所描述的技术，该文已作为参考包括在本
25

文中。

虽然在电视会议系统中跟踪发言人的传统技术对于许多应用还能令人满意，但它们也受许多限制，如果能加以克服，则可大大扩展这种电视会议系统的利用和性能。具体地说，传统的电视会议系统本质上是反应式的。因此，只有当事件已经发生，注意力才集中在该事件上。例如，一旦另外一个人发言，摄像机在聚焦到他(她)身上之前总会有一点延迟。这样远距的用户就感受不到他们似乎是在同一间屋子里进行很自然的面对面的互动。

在面对面的互动中，人们观察到，当一个人马上就要发言或要接替另一个发言人时，总会出现一些信号。例如可参阅 S.Duncan 和 Niederehe 的“该你发言时的信号”，实验社会心理学杂志，Vol.23 (2)，pp234-247 (1972)；以及 S.Duncan 和 D.W.Fiske，“面对面的互动”，Lawrence Erlbaum Publishers，Hillsdale，New Jersey，(1977)。例如，当一个人要接替另一人发言时，可以观察到微妙的线索，例如下一个发言人会身体前倾，目光注视当前的发言人或用手臂作手势。

因此在试图建立人机间自然的语言交流时，研究人员已经意识到了一个人在将不同类型的感官信息（线索）与上下文信息和以前获得的知识结合起来的能力方面的复杂程度。需要一种改进的预测事件的技术来将这种线索应用于电视会议系统中。还需要一种方法和装置来分析某些线索，例如面部表情，注视目光和身体姿势，以预测下一个发言人或其他事件。也需要一个发言人探测系统，它能将多种线索结合起来预测谁会是下一个发言者。还需要一种利用每个与会者的特征信息识别与会者发言之前会表现出那些线索来探测发言人的方法和装置。

发明概述

总地来说，本文公开了视频系统中预测事件的方法和装置。具

体地说，本发明处理音频和视频信息（或二者）来识别某一事件发生前，一个人常表现出来的一个或多个线索，包括：（1）声音线索，例如语调类型，音调与音量；（2）视觉线索，例如注视目光，面部姿势，身体姿势，手势，以及面部表情；以及（3）上述线索的组合。5 例如，在电视会议快要换发言人时，某与会者在他（她）发言之前或当现在的发言人就要结束发言时，就会显露出某些声音或视觉的线索。用这种方式，本发明就可使视频处理系统预测事件，例如识别下一个发言人。

自适应位置定位器用一种已知的方式处理音频和视频信息来确定某人的位置。另外，本发明提供一种预测发言人识别器，它识别10 一个或多个音频和视频线索从而确定下一个发言人。该预测发言人识别器接收和处理音频和视频信号，以及面部的识别分析结果，来识别一个或多个声音和视觉线索从而确定下一个发言人。预测发言人识别器产生的发言人预测结果被用来使摄像机聚焦，并获得预测15 发言人的图像。

预测发言人识别器以一种学习模式工作，根据在有或没有一个或多个预定义的声音和视觉线索的情况下与会者“会发言”或“不会发言”的概念来学习每个与会者的特征信息。然后预测发言人识别器以预测模式将特征信息中嵌入的学习特征与音频和视频信息加以比较，从而预测下一个发言人。20

参考以下详细说明和附图，就可获得对本发明，以及对本发明的特性和优点，更完全的理解。

附图简要说明

- 25 图1是根据本发明一个说明性实施例的视频处理系统的方框图；
图2是图解说明在图1的系统中实现的自适应跟踪视频处理操作的功能框图；
图3是图解说明图1的自适应位置定位器的功能框图；

图 4 是从过程角度说明图 3 的预测发言人识别器的流程图；
图 5 是说明图 4 的示范的特征简介的表格。

发明的详细说明

5 图 1 示出按本发明的视频处理系统 10。本发明处理音频和视频信息来识别某一事件发生前，一个人常表现出来的一个或多个线索，包括：（1）声音线索，例如语调类型，音调与音量；（2）视觉线索，例如注视目光，面部姿势，身体姿势，手势，以及面部表情；或（3）上述线索的组合。例如，电视会议与会者开始或结束发言之前。虽然本发明是以电视会议系统探测发言人的变化这一内容来说明，但本发明可适用于检测与人们表现出的声音和视觉线索具有关联的任何事件，这对于本专业的技术人员而言，根据此文的内容，是显而易见的。

10 如图 1 所示，系统 10 包括处理器 12，存储器 14，输入/输出（I/O）装置 15，以及自适应位置定位器 300，以下结合图 3 进一步讨论，所有这些都通过总线 17 相互连接进行通信。系统 10 还包括摇俯变焦摄像机 18，它连接到自适应位置定位器 300，如图所示。系统 10 中还可另外包括一个或多个广角摄像机（图 1 中未示出），以下结合图 3 进一步讨论，以捕捉每个与会者 22-1 到 22-N 的视觉线索。例如，可以设置一个摄像机来获得与会者 22-N 的透视图，以检测他的前倾。

15 在所示实施例中，在桌子 20 边坐了与会者 22-1 到 22-N 的电视会议应用中采用了 PTZ 摄像机 18。工作时，PTZ 摄像机 18，按照自适应位置定位器 300 根据从处理器 12 接收的指令的指引，跟踪一个关心的对象，在此例中即与会者 22-k。另外，如图 1 所示，视频处理系统 10 包括话筒阵列 16，以已知方式捕捉音频信息。

20 虽然本发明是以电视会议应用这一内容来说明的，但是，显然，视频处理系统 10 可以用在需要预测下一个发言人的其他应用中。而

且，本发明也可应用在其他类型的电视会议应用中，比如，涉及会议式座位安排的应用，以及圆桌或长方桌安排等。一般来说，系统 10 的部分 24 可用在任何应用中，这些应用可以从本文的自适应位置定位器 300 提供的改进的跟踪功能获益。因此利用系统 10 的部分 24，
5 系统 10 的部分 26 就可以用，比如，其他的电视会议安排，或其他任何需跟踪一个或多个发言人的安排来代替。显然，也可不用 PTZ 摄像机而用其他图像捕捉装置来使用本发明。此处的术语“摄像机”应包括能与本文的自适应位置定位器 300 结合使用的任何类型的图像捕捉装置。

10 应当指出，系统 10 的元件或元件组也可以代表常规台式或便携式电脑的对应元件，以及这些和其他处理装置的部分或组合。而且，在本发明的其他实施例中，处理器 12 或 PTZ 摄像机 18 的部分或全部功能和附加的广角摄像机（图 1 中未示出）或系统 10 的其他元件都可以组合成一个单一装置。例如，PTZ 摄像机 18 和附加的广角摄像机的功能可以结合成单一广角摄像机，利用广角图像的图像处理技术获得需要的近距图像。
15

另外，系统 10 的一个或多个元件可以用装入电脑，电视，机顶盒或其他处理装置中的专用集成电路（ASIC）或电路卡来实现。此处的术语“处理器”应包括微处理器，中央处理单元，微控制器，
20 或其他可以用在既定数据处理装置中的任何其他数据处理元件。此外，应当指出，存储器 14 可代表任何电子存储器，光盘或磁盘存储器，磁带存储器，以及这些和其他类型存储装置的部分或组合。

自适应位置跟踪术语

25 图 2 是说明图 1 的自适应位置定位器 300 实现的跟踪和变焦特性的功能框图。如图 2 所示，跟踪和变焦特性包括检测和跟踪操作 32 和光学变焦操作 34。参阅图像 40, 42, 44 来说明这些操作，这些图像是在系统 10 的部分 26 中为示范的电视会议应用所产生的图像。

操作 32 和 34 可以在系统 10 中，利用存储在存储器 14 或通过 I/O 装置 15 从本地或远程的存储装置接入的一种或多种软件程序，由处理器 12 和自适应位置定位器 300 完成。

工作时，PTZ 摄像机 18 产生图像 40，它包括关心的对象，例如电视会议与会者 22-k，以及另一对象，例如在关心的对象邻近的与会者 22-k+1。图像 40 作为视频输入提供给检测和跟踪操作 32，它利用已知的常规检测和跟踪技术检测和跟踪关心的对象。

例如，在电视会议应用中，该关心的对象 22-k 可能对应于当前的发言人。此时，检测和跟踪操作 32 利用音频定位来检测和跟踪该关心的对象 22-k 以确定哪一个与会者是当前发言人，以下要结合图 3 加以讨论。在另一变型中，也可利用动作检测，手势，摇头，以特殊方式动作或以特殊方式发言等来识别当前发言人。检测和跟踪操作 32 的输出包括识别具体关心的对象 22-k 的信息，以在图像 42 中标以阴影的形式表示。

图 2 的光学变焦操作 34 提供足够的变焦量以确保既可获得需要的输出图像质量，同时又允许关心的对象有一定的移动量。光学变焦操作 34 包括通过全景和倾斜操作调节关心的对象 22-k 的画面位置的画面位置调节部分和持续利用变焦操作直到满足指定的停止准则的变焦部分。通常，有多种不同类型的停止准则可以采用。在用固定的停止准则方法时，光学变焦继续进行直到关心的对象占据了图像的一定的百分比。例如，在电视会议应用中，光学变焦会继续进行，直到当前发言人的头部占据了图像垂直大小的大约 25% 到 35% 之间为止。当然，所用的具体百分比会根据跟踪应用的不同而变化。各具体应用的特定百分比可由本专业的技术人员直接确定。

如图 2 所示，光学变焦操作 34 的输出是一个光学变焦的图像 44，其中关心的对象 22-k 近似在图像的中心，占据了按上述准则确定的所需的图像百分比。图像 44 可以由系统 10 存储，例如存储在存储器 14 中，或呈现给用户。

自适应位置定位器

图 3 是图 1 的系统 10 中采用的自适应位置定位器 300 的功能框图。通常，自适应位置定位器 300 处理音频和视频信息来确定发言人的位置。关于自适应位置定位器 300 的更详细的讨论，请参阅 2000
5 年 5 月 3 日登记的美国专利申请，其申请号为 09/564016、题目为“在电视会议和其他应用中自适应位置确定的方法和装置”（Attorney Docket No.700983），该申请已转让给本发明的受让人并作为参考包括在本文中。

此外，根据本发明，自适应位置定位器 300 包括预测发言人识别器 400（以下会结合图 4 进一步讨论），用以识别一个或多个声音和视觉线索并据此预测下一个发言人。最初，在系统启动时，广角摄像机 305 和话筒阵列 16 都工作。广角摄像机 305 和话筒阵列 16
10 产生的信号可以选择在阶段 315 打上时间戳，以使自适应位置定位器 300 能判断何时产生的信号。如图 3 所示，广角摄像机 305 产生的时间戳信号被传送到面部识别模块 320。面部识别模块 320 包括一个面部检测器，它可确定某一既定的关心部分（窗口）是否可以标记为面部区域。
15 面部识别模块 320 对一既定的面部分配一个唯一的标识符。

广角摄像机 305 产生的图像、以及面部识别的结果和它们的位置都存储在画面缓冲器 325 中。但是，如果面部识别模块 320 不能对一既定的面部分配一个唯一的标识符，例如，由于发言人和广角摄像机 305 间的距离问题，那么只有面部检测信息和检测的面部在
20 图像中的相应位置存储在画面缓冲器 325 中。一些附加的信息，例如与会者的服装颜色，也可存储在画面缓冲器 325 中。服装颜色特别有用，例如，如果面部识别模块 320 不能对一既定的面部分配一个唯一的标识符，但当第一与会者离开会议室而另一与会者坐在同一位置时，面部检测仍可继续进行。
25

面部识别模块 320 可以利用，例如，美国专利申请，申请号

09/449250, 1999 年 11 月 24 日登记、题目为“在电视会议和其他应用中检测移动对象的方法和装置”，以及美国专利申请，申请号 09/548734, 2000 年 4 月 13 日登记、题目为“在电视会议和其他应用中利用组合的视频和音频信息跟踪移动对象的方法和装置”
5 (Attorney Docket No.700966) 所说明的视频定位系统来实现。此二专利已转让给本发明的受让人，并作为参考包括在本文中。如上述结合图 2 进行的讨论，视频系统也尽量聚焦（变焦）到面部，使得面部处于正确的显示纵横比的条件下。

同理，如图 3 所示，话筒阵列 16 产生的时间戳信号传送到发言人识别模块 330 和音频定位器 360。话筒阵列 16 产生的音频信号和发言人识别的结果都存储在画面缓冲器 325 中。此外，音频定位器 360 获得传送到空间变换模块 370 的识别与噪声源关联的全景（水平）和倾斜（垂直）角度的方向信息。音频定位器 310 可以利用例如在美国专利申请、申请号 09/548734、2000 年 4 月 13 日登记、题目为“在电视会议和其他应用中利用组合的视频和音频信息跟踪移动对象的方法和装置” (Attorney Docket No.700966) 以及美国专利申请、申请号 09/436193、1999 年 11 月 8 日登记、题目为“改进的信号定位装置” 中所说明的音频定位系统来实现，此二专利已转让给本发明的受让人，并作为参考包括在本文中。
10
15

在预定的时间间隔内（例如两秒）积累视频和音频信号，以便收集对应于有意义的事件的数据。在此预定时间间隔中产生的视频画面由动作检测器 350 作相互比较以检测动作。例如，如果一个与会者的手在移动，动作检测器 350 靠比较连续的视频画面检测到此动作，然后手移动的识别位置被传送到空间变换模块 370。
20

动作检测器模块 350 可任选地利用动作试探法 340 来仅仅识别具有显著的移动量的画面部分。这样，动作检测器模块 350 仅把这种滤波后的信息传送给空间变换模块 370。例如，为了检测头的转动，相应的动作试探法能指示需要转动多少才能触发响应。通常，动作
25

试探法 340 尽量使摄像机 18 聚焦在当前发言人上，而不管其他的噪声或发言人的动作。换句话说，动作试探法 340 试图识别并抑制动作检测器 350 产生的虚假事件。至于动作试探法 340 中采用的各种策略的详细讨论，请参阅，例如，Ramesh Jain 等人的“机器视力”，
5 McGraw-Hill, New York (1995)，作为参考包括在本文中。

于是，空间变换器 370 接收来自动作检测器模块 350 的位置信息和来自音频定位器 360 的方向信息。然后空间变换器 370 以已知方式映像位置信息和方向信息，以计算用来使 PTZ 摄像机 18 聚焦的边框。

10

处理声音和视觉线索

如图 3 所示，PTZ 摄像机 18 产生的视频图像，以及画面缓冲器 325 和发言人识别模块 330 中的内容都加到预测发言人识别器 400 上，以下要结合图 4 进一步讨论。还有，画面缓冲器 325 的内容包括广角摄像机 305 产生的广角图像和相应的面部识别结果，以及话筒阵列 16 产生的音频信息和相应的发言人识别结果。这样，预测发言人识别器 400 就可从广角图像和音频信息中识别每个未发言与会者 22-N 的声音和视觉线索。
15

通常，自适应位置定位器 300 按上述方式处理音频和视频信息来确定发言人的位置。如图 3 所示，自适应位置定位器 300 与预测发言人识别器 400 相互配合，按本发明预测下一个发言人的身份，以下要结合图 4 进行讨论。如图 3 所示，预测发言人识别器 400 接收来自画面缓冲器 325、PTZ 摄像机 18 和发言人识别模块 330 的音频和视频信息。预测发言人识别器 400 处理接收的音频和视频信息以识别一个或多个声音和视觉线索并据此预测下一个发言人。如图 3 所示，预测发言人识别器 400 产生的发言人预测用来聚焦 PTZ 摄像机 18。
25

图 4 是图 1 系统 10 采用的预测发言人识别器 400 的功能框图。

如图 4 所示预测发言人识别器 400 包括视频线索识别器 410 和音频线索识别器 420。当既定与会者出现在图像中时，其视频信号由视频线索识别器 410 处理，以识别一个与会者在发言之前通常表现出的一个或多个预定视觉线索，例如面部姿势（头的方向），注视目光（眼睛方向），面部表情，手和身体的姿势等。同理，音频信号由音频线索识别器 420 处理，以识别一个与会者在发言之前通常表现出的一个或多个预定音频线索，例如语调类型，音调与音量，发言速度，发言人识别和话音识别。可以用来识别下一个发言人身份的具体的声音和视觉线索处理过程在以下题目为“声音和视觉线索”一节中讨论。

10

学习模式

预测发言人识别器 400 采用一个学习模块 450，以学习模式、根据在有或没有一个或多个预定义的声音和视觉线索的情况下与会者“会发言”或“不会发言”的概念来学习每个与会者的特征信息 500。

15

如以下结合图 5 进行的讨论，每个与会者的声音和视觉线索可以存储在特征信息 500 中作为属性值的记录。此外，例如从发言人识别，面部识别或其他信息中可获得的与会者的身份也用属性值记录在特征信息 500 中。

20

记录中的每个属性可以有许多分立的或象征的值。例如，对于姿势模块，既定的与会者可能会用一些特定的姿势，例如举手要求允许发言，表示想要下一个发言。这些特定的姿势，以及其他线索模块的属性值是靠分析多个电视会议会话以确定与会者在发言前表现的手势、姿势的类型以及其它声音和视觉线索来确定的。

25

为了对与会者在可能“会发言”或“不会发言”之前通常表现出（和/或不表现）的预定声音和视觉线索进行描绘，学习模块 450 可以采用判定树（DT），例如在 J.R.Quinlan 的“学习有效分类过程及其在棋类终结游戏中的应用”，编者：R.S.Michalski 等人，在“机

器学习” (Machine Learning) 中的：人为途径， Vol.1, Morgan Kaufman Publishers Inc., Palo Alto, California (1983)；或 J.R.Quinlan “概率判定树”，编者：Y.Kodratoff 和 R.S. Michalski 等人，在“机器学习” (Machine Learning) 中的：人为途径， Vol.3, Morgan Kaufman Publishers Inc., Palo Alto, California (1990) 中所说明的判定树，均作为参考包括在本文中。在另一种途径中，可以采用 Hidden Markov 模型 (HMMs) 对与会者在可能“会发言”或“不会发言”之前通常表现出（和/或不表现）的预定声音和视觉线索进行描绘。

通常，判定树按一个训练组构建，具有节点和树叶，节点对应于需进行的某项测试，树叶对应于类别（即：“会发言”或“不会发言”）。判定树可具有的节点数取决于数据的复杂性。在最差的情况下，节点数最多可等于可能的属性值的数。举例来说，从树根到树叶的一条子通路在分解成规则时可以具有以下形式：

如果 姿势=举手， 和
15 身体姿势=前倾， 和
头部姿势=...和
面部表情=...和
注视目光=朝发言人看和
==> “会发言”

此例也在图 5 的特征信息 500 中出现。注意在上述布尔表达式中字符“？”表示“不介意”条件或通配符。

预测模式

同理，预测发言人识别器 400 采用新发言人预测器 470、以预测模式应用在特征信息 500 中的学习到的特征来预测下一个发言人。

当学习进行了一段足够的时间并建立了判定树后，在预测模式下对该判定树进行语法分析，以确定哪些模块的哪些特征足以确定谁是下一个发言人。这样，在预测模式下，新发言人预测器 470 所

用的判定树引导 PTZ 摄像机 18 并确定利用哪些模块来获得谁是下一个发言人的结论。

应当指出，在会话中预测谁是下一个发言人可以看作是一个数据开发/知识发现的问题。在此域中，目的是看能否从数据中找到一种模式。我们想要建立的具体模式就是与会者是否表现出一些线索，预示着他们可能参加对话。专门采用判定树来学习在数据中含有的一同时发生的情况与明显学到的结果之间的因果关系。例如可以学到以下的规则：如果一个与会者举手、身体前倾、而且该规则中其他同时发生的事件为未知，则该与会者可能即将发言（发生的结果）。

为了检测发言人的转换，当判定树在连续窗口给出一个不同与会者的类别（通过面部识别/发言人识别/音频定位），系统就假定不同的发言人开始发言。能用来表示当既定行为足以构成提示下一个发言人的“线索”时的精确阈值可以靠经验确定。

如前所述，图 4 的预测发言人识别器 400 利用图 5 所示的特征信息 500 对某一与会者在可能“会发言”或“不会发言”时通常表现出或不表现的一个或多个预定声音和视觉线索进行描绘。图 5 是说明特征信息 500 可能实施的示范表格。如图 5 所示，特征信息 500 包括许多记录，例如记录 505-515，当分解为规则时，每个都与从判定树树根到树叶的不同通路相关联。对于每条判定树的通路，特征信息 500 识别定义通路的字段 550-565 中的属性值对和字段 570 中的相应概念类别。

这样，当特征信息 500 中的既定规则提示一个新的与会者即将发言时，预测发言人识别器 400 可提供一个预测 PTZ 值给摄像机 18，以使摄像机 18 在该与会者一开始发言时就聚焦在预测发言人上。在一个实施例中，可用另一台 PTZ 摄像机跟踪预测发言人，当发言人开始发言时可以选择相应的图像作为系统 10 的输出。

视觉和声音线索

如前所述，视觉线索识别器 410 识别一个与会者在发言之前通常表现出的一个或多个预定视觉线索，例如手势，面部姿势，注视目光，面部表情，手和身体的姿势，可能还有感情等。例如，注视目光的信息在识别一个人的注意焦点方面起重要作用，即，看哪里，他在注意些什么。注视目光方向由两个因素决定：头的取向和眼睛的取向。头的取向决定整个的注视方向，而眼睛的方向可以决定精确的注视方向并受头的取向的限制。因此，当一个人即将发言时，他的目光通常聚焦在当前的发言人上。

同理，以下每对属性值对应于提示某人要开始发言的视觉线索：

属性	值
注视目光	眼睛看当前发言人
手势	举手或手指
面部姿势	面部对着当前发言人；点头
面部表情	微笑
身体姿势	前倾

面部表情

面部表情可以按照以下文章中说明的技术获得，例如：“应用于人机界面的连续视频（图像）的面部分析”，Ph.D.Dissertation, University of Illinois at Urbana-Champaign (1999)；或 Antonio Colmenarez 等人的“嵌入的面部和面部表情识别的概率框架”，Proc. of the Int'l Conf.on Computer Vision and Pattern Recognition, Vol.I,592-97, Fort Collins, Colorado (1999)，它们已作为参考包括在本文中。面部表情的强度可以按照以下文章中说明的技术获得，例如美国专利申请、申请号 09/705666、2000 年 11 月 3 日登记、题目为“利用双向星形拓扑 Hidden Markov 模型估计面部表情的强度”，该申请已转让给本发明的受让人，并作为参考包括在本文中。

头部姿势/面部姿势

头部或面部姿势可以按照以下文章中说明的技术获得，例如：
Egor Elagin 等的“基于群聚图匹配技术的面部自动姿势估计系统”，
Proc. of the 3rd Int'l Conf. on Automatic Face and Gesture Recognition
5 (第三届自动面部和姿势识别国际会议文集), Vol. I, 136-141, Nara,
Japan (1998.4.14-16)，已作为参考包括在本文中。

注视目光

10 注视目光以及面部姿势可以按照以下文章中说明的技术获得，
例如：John Heinzmann 和 Alexander Zelinsky 的“利用增强实时跟踪
范例作三维面部姿势和目光注视点估计”， Proc. of the 3rd Int'l Conf.
on Automatic Face and Gesture Recognition (第三届自动面部和姿势
识别国际会议文集), Vol. I, 142-147, Nara, Japan (1998.4.14-16)，已
作为参考包括在本文中。

15

手势

20 手势可以按照以下文章中说明的技术获得，例如：Ming-Hsuan
Yang 和 Narendra Ahuja 的“利用运动轨迹识别手势”， Proc. of the IEEE
Computer Society Conf. on Computer Vision and Pattern Recognition,
Vol.I, 466-472, Fort Collins, Colorado (1999.6.13-15)，已作为
参考包括在本文中。

身体姿势

25 身体姿势可以按照以下文章中说明的技术获得，例如：Romer
Rosale 和 Stan Sclaroff 的“不跟踪身体各部分理解身体姿势”， Proc.
of the IEEE Computer Society Conf. on Computer Vision and Pattern
Recognition, Vol.2, 721-727, Hilton Head Island, South Carolina
(2000.6.13-15)，已作为参考包括在本文中。

同理，音频线索识别器 420 识别一个与会者更换在发言人改变之前常表现出的一个或多个预定音频线索，例如非话语声音，比如咕隆一声或清清喉咙。音频线索可以按照以下文章中说明的技术识别，例如：Frank Dellaert 等人的“在话音中识别感情”，Proc. of Int'l Conf. on Speech and Language Processing (话音和语言处理国际会议文集) (1996)，已作为参考包括在本文中。一旦音频线索源被识别，就可利用发言人识别功能来识别谁在发言。此外，也可利用话音识别技术进一步改进发言人预测。例如，假定某人 A 在发言，正当他还 5 在讲话时某人 B 开始说：“我不同意你的意见”。如果话音识别系统已经接受过这种句子的训练，那么，在系统能识别该句子的那一刹那就暗示某人 B 可能是下一个发言人。
10

发言人的情绪可以从声音和韵律等特性来估计，例如语速，音调与音量，语调，强度等。发言人的情绪常提示发言人即将结束发言。发言人的情绪可以按照以下文章中说明的技术识别，例如：Frank 15 Dellaert 等人的“在话音中情绪”， Proc. of Int'l Conf. on Speech and Language Processing (话音和语言处理国际会议文集) (1996)，已作为参考包括在本文中。

如前所述，本发明可用来检测与人们表现出的声音和视觉线索有关联的任何事件。除了上面已充分说明的检测更换发言人之外，
20 另外的示范事件和对应线索包括：

事件	线索	动作
某人睡着了（让他睡）	头下垂，眼睛闭上，打呼	调低音乐，关掉电视，开始录制节目
某人睡着了（把他弄醒）	头下垂，眼睛闭上，打呼	调高音乐，启动闹铃
学生在课堂上想发言	举手	提醒老师
想去厕所的小孩	跳动，捂住身体	提醒家长

所以，本发明可以用来预测许多事件并在其之前采取相应的行动。例如，本发明可以用在车辆中检测司机是否快要睡着了，如果检测到此情况，就采取适当的措施。又如，本发明可以用来检测看电视的人是否睡着了，并能采取适当的措施开始录制其余的节目，以及关掉电视，电灯和其它的电器。

显然，以上说明的实施例和各种变型仅是为了说明本发明的原理，在不背离本发明的范围和精神的情况下本专业的技术人员可作各种改动。

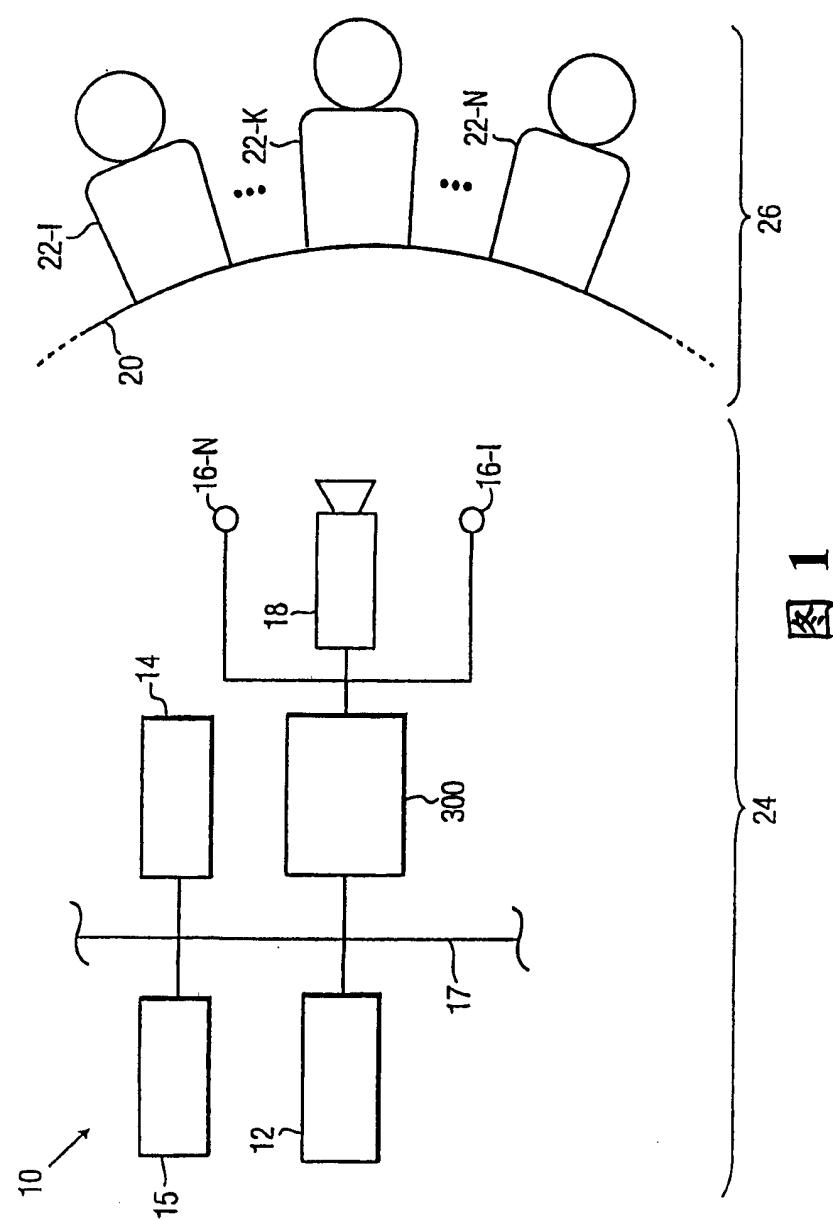


图 1

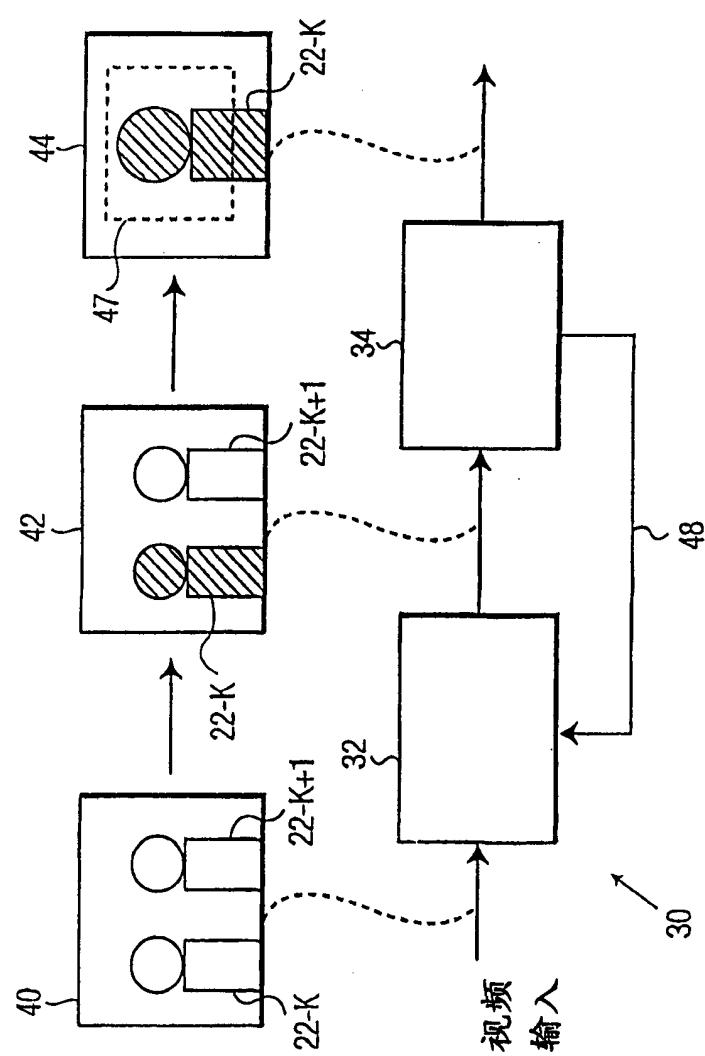


图 2

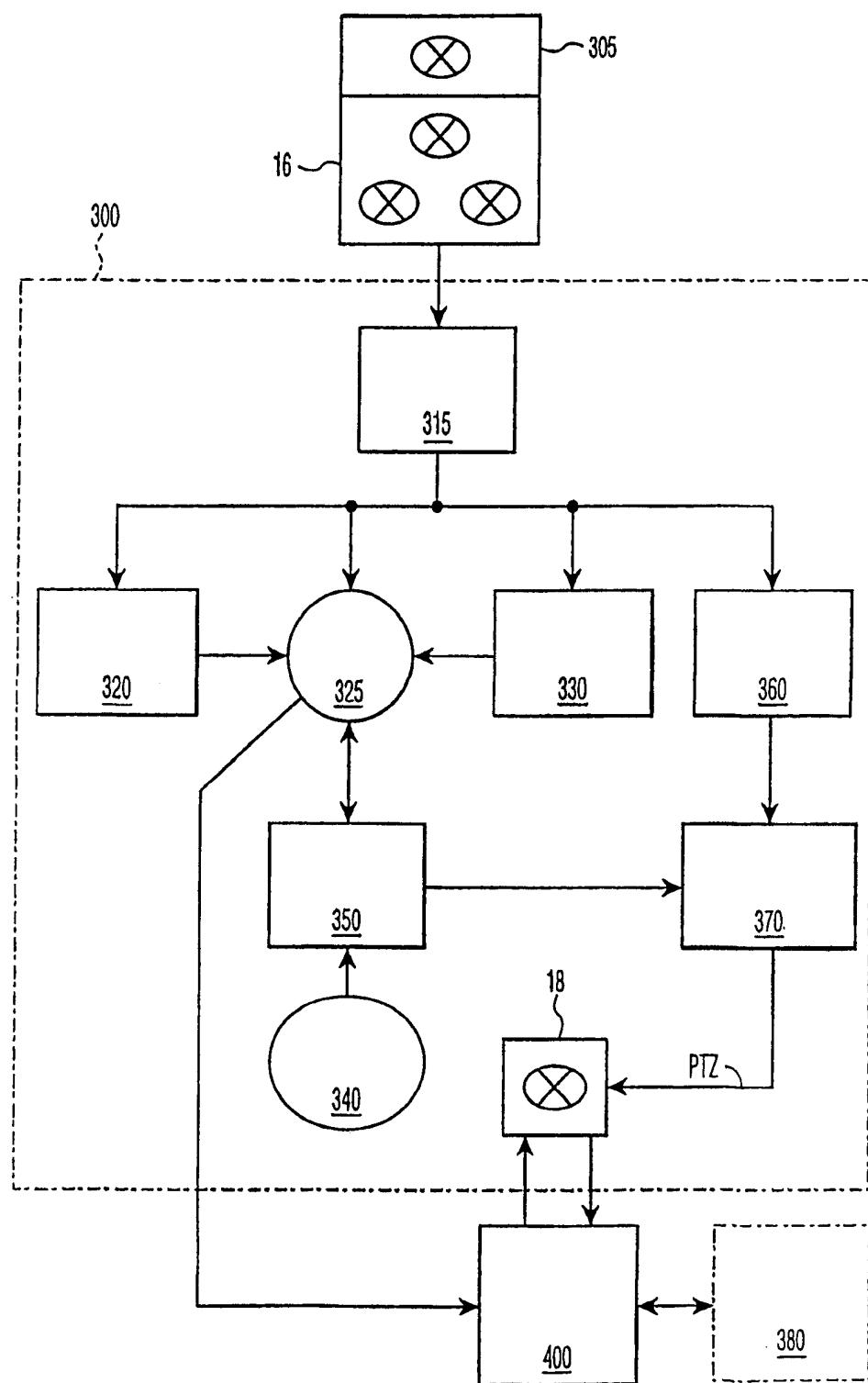


图 3

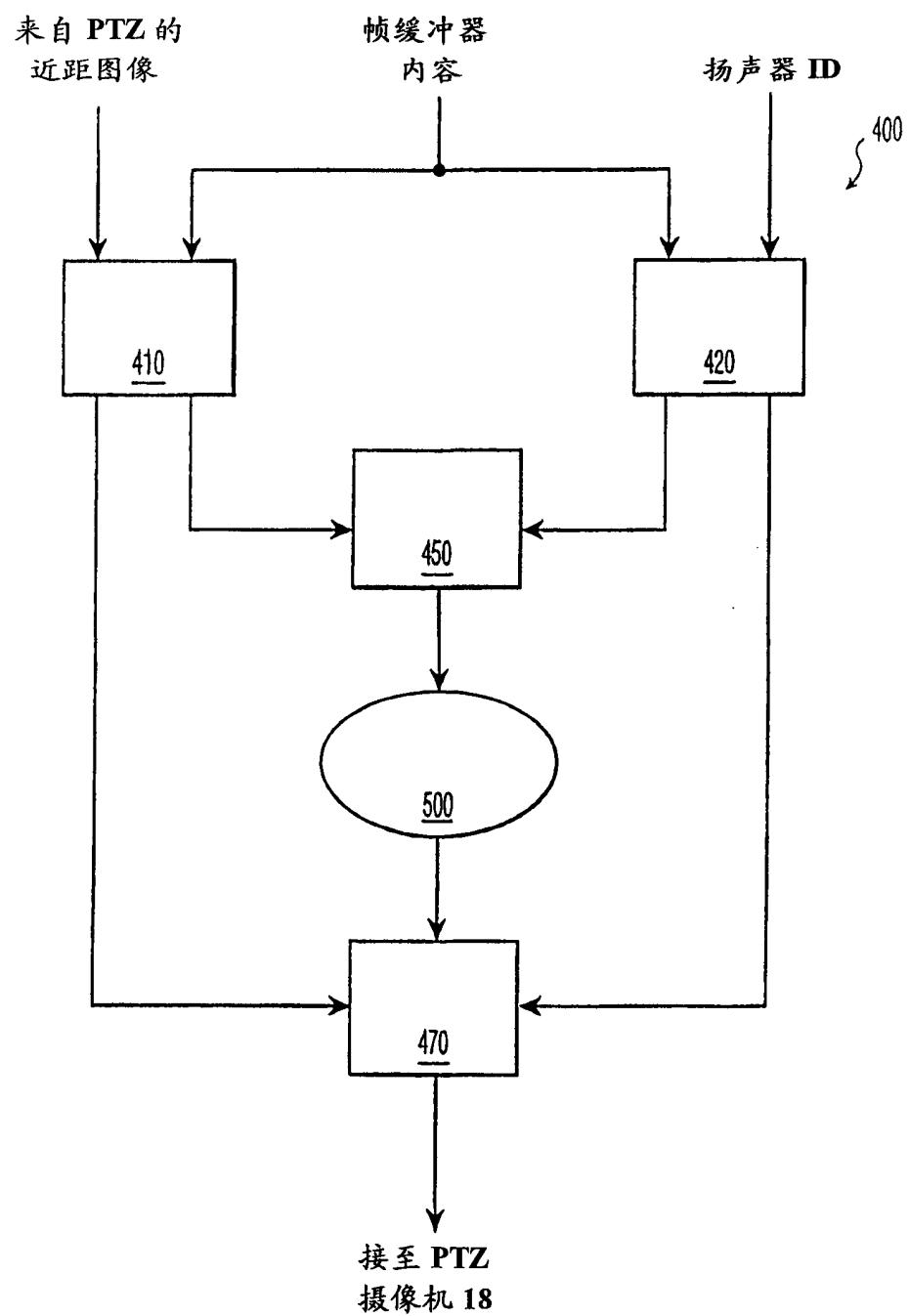


图 4

图 5

	发言人识别器 <u>550</u>	属性值对 1 <u>555</u>	属性值对 2 <u>560</u>	...	属性值对 N <u>565</u>	概念 <u>570</u>
<u>505</u>						
<u>510</u>	John Smith	手势=举手	身体姿势= 前倾		音调=?	将要发言
...						
<u>515</u>						