基于立体视觉的自然手势识别*

张 凯+, 葛文兵, 汪国平, 董士海

北京大学信息科学技术学院人机交互与多媒体实验室, 北京 100871

摘 要: 本文提出了一种基于立体视觉的自然手势识别方法。本文方法有着如下特点: i) 无穿戴自然手势交互, 在满足平面近似的条件下, 可任意扩充手势语义; ii) 模板库小, 仅仅需要存储每种手势的正面图像, 便于高效匹配; iii) 提出了一种新的立体匹配算法, 将手势的视差图简化为平面, 用以快速获取手势和手臂的三维信息; iv) 可自动调整相机位置, 得到手势的正面图像, 用于模板匹配。实验结果表明, 本文算法对各种朝向的手势都有着良好的支持, 识别率高, 强壮性好。

关键词: 手势识别;立体匹配;视差

1. 引言

在自然交互领域, 手势识别一直都是研究的热点。作为一种通用的基础算法, 手势识别广泛应用于交互游戏、三维设计、机器人研究等多个领域。利用数据手套, 人们可以获得精确的手势数据[1]。但是, 这类方法在交互上不自然的缺点是显而易见的, 因此, 在对精度要求并不十分严格的情况下, 无穿戴的视觉方法更受青睐, 这也是现在手势识别技术的主要研究方向。目前, 手势识别算法分为两大类: 基于模型的识别和基于图像的识别。基于模型的方法利用多相机重建手势的二维信息, 将重建后的体数据与给定的三维模型相匹配。这类算法精度高, 但是计算代价昂贵, 且稳定性差[2－5]。基于图像的方法利用多个角度的手势图像作为模板, 并插值生成模板库中没有的手势图像。由于无法控制手势的朝向, 这类算法往往需要庞大冗余的模板库, 为手势匹配增加了不必要的开销, 当光线条件不理想时, 也容易得到错误的匹配[6－9]。

在计算机视觉领域, 立体匹配在近十年来也一直都是研究的热点问题。给定一对水平校准的立体图像, 对场景中任一空间点 P, 考虑其在左右图中的映像 p 和 q。因为图像是校准的, 由极线几何可知, p 和 q 都在同一条水平扫描线上, 它们之间的位置差值称为视差(disparity)。立体匹配以立体图像中的某一幅图(通常是左图)为参考图, 确定每一点的视差及遮挡关系, 并得到场景视差图(disparity map)。由几何关系可推

* *本文受到国家重点基础研究发展计划 No.2004CB719403, 国家高技术研究发展计划项目 No.2004AA115120, 自然科学基金项目 No.60473100 的资助

*张凯, Email: zk@graphics.pku.edu.cn

知，视差和景物深度成反比[10，11]，求得视差之后，可以很方便的得到物体的深度信息，并重建三维模型。基于 Graph-cut 和信度扩散（belief propagation）的立体匹配算法可以达到相当高的精度[12，13]，但这类算法效率不高，不适合实时交互的需要。

本文结合前述两类手势识别技术的优点，利用立体视觉的方法求得手势的正面图像，然后与预定义的模板进行匹配，从而确定语义。本文算法基于两个假设。第一个是关于手势的“平面性假设”。我们注意到，由于手部纹理的近似性，我们可以安全的忽略手势的自遮挡现象，将手势的视差图简化为平面。实际上，这一个假设对于大多数手势都是满足的。第二个假设是手势和手臂的“共面假设”，即假定在手势运动过程中，手势和手臂基本处于同一个平面。如果我们能得到手势平面的参数和手臂的朝向信息，就能对相机位置进行刚体变换，求得手势的正面图像。

本文的后续部分如下组织：第二节介绍算法的预处理过程；第三、四节是本文的重点部分，首先介绍以场景平面模拟手势的立体匹配算法，然后依据得到的手势和手臂的三维信息，调整相机位置，求得手势正面图像，进行语义判定；最后讨论实验结果并进行小结。

2．预处理

算法的预处理过程包括三部分。首先，我们需要定义不同的手势，用以表达不同的语义。我们对每个手势正面拍照，使相机主平面与手掌平面平行，相机主平面的 Y 轴方向和手臂方向平行，相机光心与手势质心的连线与相机主轴重合。对每一个手势照片，求得其外接矩形，并将此矩形缩放至一定大小（譬如，64×64 像素），作为手势模板。由于仅仅存储手势的正面图像，我们的模板库十分简洁，也便于高效匹配。在满足平面假设的条件下，我们可以任意扩充手势模板。图 1 展示了几种不同的手势。

图 1 手势模板示例。这里列举了八种常见手势。

手势运动时，场景可以分为前后两部分：前景包括运动的手臂和手势，后景包括静止的身体和背景。为了简单起见，我们在进行匹配时仅仅考虑运动的前景，而将后景除去。除去静止后景的方法有很多种，一种方法是对多帧图像进行累加后求平均，然后与目标图相减，灰度差值大于给定阈值的就是前景。或者对特定的几帧求得密集匹配，以某一视差阈值进行筛选，视差小于该阈值的像素被认为是后景。

最后，我们对本文算法使用的两个相机进行完全标定。我们以左边相机的光心为世界坐标原点，将其投影矩阵记作：$P = K[I \mid 0]$。其中 K 是相机内参数矩阵，I 是 3×3 单位矩阵。此外，为了组成合适的立体装置，我们还需要求得两个相机的水平校准矩阵。为便于叙述，我们假定后续处理的所有立体图像都是经过水平校准的。

3．提取三维信息

这一节首先介绍本文立体匹配算法的理论基础，包括图像分割，匹配约束和代价函数，然后计算手势平面和手臂朝向。后续处理中，我们都以左图为参考图。

3．1 图像分割

在进行匹配之前，需要对前景进行图像分割。这里的分割有两个层次。第一层次的分割是为了将手与手臂分开。容易看出，手部的颜色纹理是相近的，如果手臂上有衣袖的话，那么很容易通过基于颜色的图像分割算法区分手与手臂。在没有衣袖的情况下，我们在手腕上佩戴一个标识环状物，用以指导这一层次的图像分割。

第二层次的图像分割用于立体匹配。这一思想基于如下的假定：相邻的像素，如果其灰度值相近，那么它们的视差相同。通过合适的图像分割算法（如[14]）将立体图像分为若干小块，再以每小块为匹配基元求视差，可以大大提高算法效率和匹配效果。图像分割技术作为一个有效的辅助手段，现已为许多立体匹配算法采用[12，13]。

3．2 匹配约束

我们将水平校准的左右两帧立体图像标记为 I_L，I_R，分别对应视差图 D_L，D_R。p 用来表示左图像素。在完整的表述下，像素由 x，y 两个方向的坐标确定。由于图像是校准的，所以我们仅仅考虑 x 坐标而忽略 y 坐标。为方便起见，我们直接使用 p 表示像素的 x 坐标值。$I_L(p)$分别表示像素 p 的灰度值。D_L 为 I_L 中每一个像素 p 指定视差 d：$D_L(p) = d$。由视差的定义可知，d 同时是 I_R 中像素 $p - d$ 的视差，即有：$D_R(p - d) = d$。(p, d) 称为一个匹配，也可以认为是一个三维点。公式 $D_L(p) = D_R(p - D_L(p))$一般称为像素的一致性约束。由于本文算法以图像块作为匹配基元，图像块一般涉及多个像素，所以需要单独定义图像块的一致性约束：

图像块一致性约束 记 S 中满足一致性约束的像素和所有像素的比率为一致性比率，为一致性比率定义阈值 γ。我们认为，图像块 S 满足一致性约束当且仅当 S 的一致性比率大于 γ。

满足一致性约束的图像块称为基本控制块（ground control segment），基本控制快中的像素称为基本控制点（ground control point）。

3．3 代价函数

代价函数就是灰度匹配函数，使代价函数最低的视差值被认为是最佳视差。图像块是我们算法的匹配基元，其代价函数就是将图像块中所有像素的代价值进行求和平均。对于给定的分块 S 和视差 d，$|S|$ 表示 S 中像素的数目，我们有：

$$Cost(S,d) = \frac{1}{|S|}\sum_{p \in S} |I_L(p) - I_R(p-d)| \tag{1}$$

3．4 手势平面匹配

在进行第一层次的分割后，场景由手和手臂组成。如前所述，由于手上各个部分的纹理颜色相近，且深度变化不大，我们可以忽略由手势产生的自遮挡，而简单认为手势处在某一个空间平面上。手臂由于不存在自遮挡现象，可以直接使用平面模拟。这样，需要识别的场景就由两个平面组成，视差较大的就是手势平面，视差较小的是手臂平面。

图像空间中，场景平面方程可表示为：

$$\frac{1}{Z} = d = ax + by + c \tag{2}$$

其中，Z 表示像素深度值，d 表示视差，$(x，y)$为像素坐标值。

初始时，依据代价函数(1)，计算每个图像块的最佳视差，并求得基本控制块。将基本控制块中的像素值代入(2)式，利用最小二乘法，可得到平面的初始解。应用(2)式，平面的初始解又可限定未定图像块（非基本控制块）的视差取值范围，继续求得基本控制块。如此反复迭代，最终得到手势平面的最优解。

由图像空间中的平面方程得到三维空间中的平面方程相对简单。记三维空间中手势平面为 $\pi_G = (\pi_{G1}, \pi_{G2}, \pi_{G3}, \pi_{G4})^T$。在透视几何中，平面上的任一点$(X, Y, Z, 1)$和其图像平面坐标$(x，y)$存在如下比例关系：

$$\frac{f}{Z} = \frac{x}{X} = \frac{y}{Y} \tag{3}$$

其中 f 是相机焦距，我们假定为已知。将(2)、(3)代入平面方程，可得：

$$\pi_{G1} \cdot x + \pi_{G2} \cdot y + \pi_{G3} \cdot f + \pi_{G4} \cdot d \cdot f = 0 \tag{4}$$

利用最小二乘法，可以完全求得手势平面 π_G。

后续处理还需要知道手势的质心。根据手势质心的图像坐标，由(2)、(3)，可以很方便的求得手势质心的空间位置 C_G。

3．5 手臂朝向估计

和手势平面一样，手臂的朝向也是为后续的相机位置调整服务的。在这里，我们把手臂看作图像空间中的直线 L：

$$ex + fy = g \tag{5}$$

$(x，y)$为像素坐标值。以手臂平面上每个像素为输入参数，直线 $L = (e，f，g)^T$ 可以用最小二乘法求得。为便于后续处理，我们改变 g 值，限定 L 通过手势的质心。若当前相机的投影矩阵为 P，映射到直线 L 的场景平面 $\pi_D = (\pi_{D1}, \pi_{D2}, \pi_{D3}, \pi_{D4})^T$ 为：

$$\pi_D = P^T \begin{bmatrix} e & f & g \end{bmatrix}^T \tag{6}$$

4．判定手势语义

手势语义通过手势图像与手势模板比较得到。由于所有手势模板都使用正面图像，因此，需要利用手势和手臂的三维信息对帧图像进行变换，得到正面的手势图像。

4．1 相机位置调整

为得到正面手势图像，需要对相机和图像进行透视变换。记相机的成像中心为光心，过光心且垂直于成像平面的直线称为相机主轴，过光心且垂直于主轴的平面称为相机主平面，易知主平面与成像平面平行。我们对相机位置的调整分为三步进行，限于篇幅，下文省去了变换矩阵 R_1，T_2，R_3 的表达式，感兴趣的读者可以参见[15]。

首先固定光心，对相机施以旋转变换 R_1，使得相机主平面与手势平面 π_G 平行。然后对相机应用平移变换 T_2，使得光心与手势质心 C_G 的连线和相机主轴重合。变换后的投影矩阵和图像分别为：

$$P' = K[I \mid 0]T_2^{-1}R_1^{-1}, \quad x = P'X \tag{7}$$

其中 X 是空间点坐标，可由(2)、(3)式求得。x 是图像点坐标。对变换后图像空间上不存在的像素，采用插值的方法计算。

接下来应用(6)式求得手臂朝向在图像空间上的直线 L 以及 L 和相机光心所确定的场景平面 π_D。由于 L 通过手势质心，那么 π_D 必然通过相机主轴，从而垂直于相机主平面。为使相机 Y 轴与手臂的方向平行，只需要使其落在平面 π_D 上，这可以通过绕主轴的旋转变换 R_3 完成。此时，投影矩阵和图像变换为：

$$P'' = K[I \mid 0]R_3^{-1}T_2^{-1}R_1^{-1}, \quad x = P''X \tag{8}$$

这样，我们就得到了手势的正面图像。图 2 给出了调整相机位置的一个例子。

(a)　(b)　(c)　(d)

图 2 相机位置调整。(a)是初始图像；应用旋转变换 R_1 后，相机主平面与手势平面平行，得到图像(b)；应用平移变换 T_2 后，相机光心和手势质心的连线与相机主轴重合，得到图像(c)；应用旋转变换 R_3 后，相机 Y 轴与手臂朝向平行，得到手势正面图像(d)。

4．2 模板匹配

模板库中的每个手势都定义有不同的语义。有了手势的正面图像，语义判定就相

对简单。记待比较的图像为 I，模板库为$\{T\}$。将 I 缩放至模板大小，通过计算匹配误差，可找到最佳匹配，如(9)式所示：

$$T_{prop} = \arg\min_{T\in\{T\}}\left(\sum_{p}|I(p)-T(p)|\right) \tag{9}$$

5．实验结果与分析

本系统针对图1给出的八种手势进行了识别，每种手势都取三种以上的朝向，在三种不同的图像分辨率下进行实验。实验结果表明，除了少数极端情况（如手势与相机平面垂直，导致无法确定手势平面参数）外，我们的算法都能做到正确的识别。图 3 给出了四种手势在不同姿态下的识别例子。

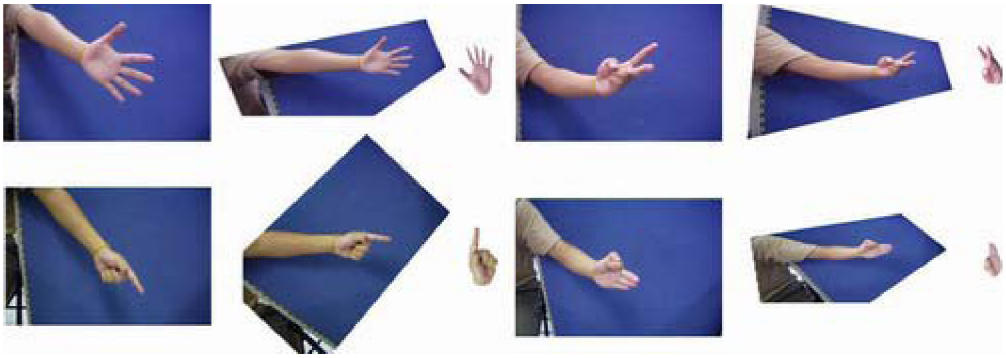

图 3 手势识别示例。图中给出了四种手势的识别过程，对每种手势，分别给出了手势的原始图像，算法求得的手势正面图像和识别结果。

第一节中，我们给出了两个算法假设。在满足第一个平面性假设的条件下，我们可以任意扩展手势模板，扩充手势语义。第二个共面假设并不严格，当手臂和手势不共面，但两者平面的夹角较小时，我们仍能给出正确的匹配。如图 3 中的“V”型手势，手臂和手势有着明显的夹角，但这并没有影响到匹配结果。

6．小结

本文提出了一种基于立体视觉的自然手势识别方法。利用平面立体匹配算法，我们能够很方便的得到手势的平面参数。结合手臂的朝向信息，算法可以自动调整相机位置，得到手势的正面图像，应用于模板匹配，确定手势语义。在满足平面假设的条件下，可以任意扩展手势模板，扩充手势语义。实验结果表明，本文算法识别率高，强壮性好，对各种朝向的手势都有着良好的支持。

参考文献

[1] T. Sarner, J. Weaver, and A. Pentland. "Real-time American Sign Language recognition using desk and wearable computer based video". IEEE Trans. Pattern Anal. Mach. Intel. 1998. vol.20, 1371-1375

[2] N. Jojic, et al. "Detection and estimation of pointing gestures in dense disparity maps". Proc. of the 4th Intl. Conf. on Automatic Face and Gesture Recognition. 2000. Grenoble

[3] Y. Sakagami et al. "Intelligent ASIMO: System overview and integration". IEEE Intelligent Robots and Systems. 2002. Genova. 2478-2483

[4] J.M. Rehg and T. Kanade. "Visual tracking of high dof articulated structures: An application to human hand tracking". In Eur opeanConfer ence on Computer Vision. 1994. 35-46

[5] J. Lee and T. L. Kunii. "Model-based analysis of hand posture". IEEE Computer Graphics & Applications. September 1995. vol.15(no.5), 77-86

[6] L. Bretzner, I. Laptev, and T. Lindeberg. "Hand gesture recognition using multi-scale colour features, hierarchical modesl and particle filtering". Proc. of the 5th Intl. Conf. on Automatic Face and Gesture Recognition. May 2002. Washington D.C. 423-428

[7] V. Pavlovic, et al. "Visual interpretation of hand gestures for human-computer interaction: a review". IEEE Trans. on Pattern Anal. Mach. Intel. 1997. vol.19(no.7), 677-695

[8] Y. Zhu, G. Xu, and D.J. Kriegman. "A real-time approach to the spotting, representation, and recognition of hand gestures for human-computer interaction". Computer Vision and Image Understanding. 2002. vol.85(no.3), 189-208

[9] X. Zhu, J. Yang, and A. Waibel. "Segmenting hands of arbitrary color". Proc. of the 4th Intl. Conf. on Automatic Face and Gesture Recognition. March 2000. Grenoble. 446-453

[10] D. Scharstein and R. Szeliski. "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms". International Journal of Computer Vision. 2002. vol.47 (no.1), 7-42

[11] M. Z. Brown, D. Burschka, and G. D. Hager. "Advances in Computational Stereo". IEEE Transactions on Pattern Analysis and Machine Intelligence. August 2003. vol.25(no.8), 993-1008

[12] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. "Stereo Matching using Belief Propagation". IEEE Transactions on Pattern Analysis and Machine Intelligence. 2003. vol.25(no.7), 787-800

[13] Li Hong and George Chen. “Segment-based Stereo Matching using Graph Cuts”. In IEEE Computer Vision and Pattern Recognition. 2004. 74-81

[14] D. Comaniciu and P. Meer. "Mean shift: A Robust Approach toward Feature Space Analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. vol.24(no.5), 603–619

[15] R. Hartley and A. Zisserman. "Multiple View Geometry in Computer Vision". Cambridge: UK, Cambridge Univ. Press. 2000

Stereo Vision Based Natural Gesture Recognition

Zhang Kai +, Ge Wenbing, Wang Guoping, Dong Shihai

(HCI & Multimedia Lab, School of Electronics Engineering and Computer Science, Peking University, 100871, China)

+Corresponding author: Zhang Kai. Phn: +86-10-62765819-813. E-mail: zk@graphics.pku.edu.cn

Key words: gesture recognition; stereo matching; disparity

Abstract: Gesture recognition is one of the most active research area in human-computer interface, and stereo matching for extraction of three-dimensional scene structure has also been an intense area of research for decades. In this paper we combine the techniques of two aspects and propose a stereo vision based hand gesture recognition algorithm by template matching. Our algorithm is mainly based on two assumptions. The first is gesture's planar hypothesis by which we neglect the self-occlusion of hand gesture and deem the gesture's disparity map as a plane. The second is conplane hypothesis by which we presume hand and arm are in the same plane. If we have got the gesture plane and the arm direction, we can rotate and translate the camera to get the gesture's frontal image, then gesture semantic can be recognised by comparing the frontal image with each template image. There are four contributions of our paper. First, we present a wearless gesture recognition method and natural gestures are expandable. Next, we use a compact gesture template for high efficient recognition, including only gestures' frontal images. As a third contribution, we use a plane-based stereo matching algorithm to get the 3D geometry infomation of hand plane and arm direction. Finally, gesture's frontal image is retrieved by automaticly rotating and translating camera's position. Experiments demonstrate that high recognition rates are achieved in various types and poses of gestures, which shows our method is strong and promising.