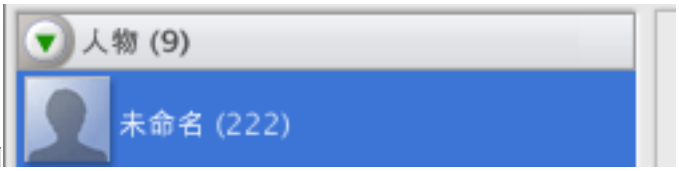
Grouping 流程

作者：崔逸卿

1. 人工初筛。人工筛选掉一些使用价值不高的图。例如：1. 没有人脸的照片。2.没有主题的照片。为拍摄者随手拍，画面中没有主要人物，或者一看就是拍了一堆路人，不包含主要人物的图片。3. 画面主题不是人的图片。例如图片主要拍摄花花草草，却在画面的角落里乱入了几个人脸。4. 无意义的人脸。比如拍摄的电视机、电影、海报、广告中出新的人脸。5. 质量不高的人脸。比如图片中的人物太模糊、90°侧脸、脸被遮挡的面积过大（例如超过 40%）等。
2. 将图片导入 picasa。简单起见，我们在桌面新建一个文件夹，将完成步骤 2 之后的图片文件夹全部放进去。然后打开 picasa，设置扫描图片路径。具体：工具 -> 文件夹管理器，打开画面如下：

为了仅导入我们需要的图片，此处将所有的其他文件夹全部选择红色的叉叉(从 picasa 中删除)，仅仅对之前新建的图片文件夹选择“总是扫描”，然后确定。此时 picasa 会开始扫描图片并检测人脸（如果左部浏览栏中有无关文件夹，可以手动选择，右键“从 picasa 中删除”）。

3. 人工聚类。虽然工具会协助我们聚类一部分人脸，但其中有一定比例的错误，需要人工参与工作。具体方法：点击工具左边的“人物”标签下的“未命名”，例如 

，然后在右边选择“展

开群组”，例如

（如果没有展开群组，只有“按头像分组”和“显示忽略的头像”，则说明已经展开了），此时会发现所有的人脸已经按照一定的顺序归类在一起了。此时随便选择一张图片，为他命名并创建联系人，则所有的被软件识别为同一个人的人脸就自动归到了这个名字下。有许多软件不能确认的图片需要人手工归类。注意：不同的人要用不同的名字，名字只使用字母和数字，但一定不要使用纯数字命名，不能重复。对于那些出现频率很少的“路人”，可以选择忽略他，不为他归类。最终我们需要的结果是每一个主要人物的头像都被聚类到了一起，并且有一个独一无二的名字。

4. 根据数据库生成 xml。进入到工具 exportpicasa-0.4.1-win32 目录下，运行工具 exportpicasa.exe。直接运行 exportpicasa.exe 可以得到工具的使用方式。以我的电脑 win7 为例，运行的命令为 exportpicasa.exe –d “C:\Users\username\AppData\Local\Google\Picasa2” –o “C:\Users\username\Desktop\exportpicasa-0.4.1-win32”，其中-d 后的参数为数据库文件所在位置（AppData 可能为隐藏目录，因此可能需要设置查看所有文件夹），-o 后的参数表示生成的 index.xml 文件存放的目录。
5. 根据 index.xml 生成结果表格。将生成的 index.xml 移动到和脚本 processXml.py 同样的目录下，然后运行 processXml.py 脚本。程序会要求输入结果文件前缀，输入源文件夹名字即可。比如这个 index.xml 是根据图片文件夹 renren_123456 中的图片聚类生成的，则输入 renren_123456。程序会输出对应的 csv 文件。这就是我们最终需要的文件。这个文件的格式如下图：

	A	B	C	D	E	F	G	H	I
1	picName	personName	faceId	left	top	right	bottom		
2	P1000031.JPG	tr	0b842a02-fa08-4218-bc68-1b801a4c39f3	0.358724	0.394095	0.414923	0.483223		
3	P1000031.JPG	m1	0b286f8f-34ee-430d-97dc-78c65b4300a0	0.233066	0.416663	0.293843	0.513313		
4	P1000031.JPG	lj	4bbe5fbc-a696-4712-a52b-5d86b6ebb063	0.128695	0.552377	0.212451	0.686625		
5	P1000031.JPG	ld	9ddf0734-caf1-4365-b06d-3b2d0f0b62cc	0.74197	0.44242	0.811856	0.554116		
6	P1000032.JPG	tr	c92bbfcd-64b1-417c-9e38-85c4dcee3b2b	0.728084	0.390051	0.786892	0.483223		
7	P1000032.JPG	ll	3ef6809f-0ec8-44a8-b3f1-ce58d103440b	0.249348	0.364294	0.326818	0.488426		
8	P1000032.JPG	lj	5f7e124e-bec4-4c29-a123-6d07c2bc728b	0.440101	0.484947	0.489372	0.564233		
9	P1000035.JPG	tr	2acdcb5f9-586c-4843-b581-aa1eef9ece10	0.110674	0.405966	0.176867	0.511574		
10	P1000035.JPG	m1	cdba77ce-2a8e-424f-b3cd-cf1556471d6f	0	0.437507	0.0668345	0.554971		
11	P1000035.JPG	xb	6c547aa3-8cab-4fb3-b650-cdfae0b498d6	0.403647	0.449943	0.450523	0.52488		
12	P1000036.JPG	m1	599065bf-830c-4453-97fb-82ee93380269	0.4757	0.44329	0.520623	0.515908		
13	P1000036.JPG	tr	93046cd7-b0c7-4472-8765-b1100ea7433e	0.579431	0.427375	0.611322	0.478294		
14	P1000036.JPG	ll	07ae1089-644c-49ca-b5d7-0fd88c7f445a	0.334417	0.350118	0.471794	0.570016		
15	P1000036.JPG	xb	16f2de22-2012-4828-ba13-2b4cda51f955	0.815106	0.425055	0.897566	0.55671		
16	P1000040.JPG	ll	64bbd8d2-af2d-405c-83a6-7cea4a6eb054	0.353079	0.359091	0.508034	0.606775		
17	P1000040.JPG	m1	efb30860-96b3-4e2c-b704-f91ec0114fd3	0.930556	0.466728	0.999344	0.580728		
18	P1000040.JPG	lj	1c5ae691-10af-4d28-8a6c-e84611a5f741	0.768658	0.643229	0.820523	0.725689		
19	P1000041.JPG	ll	fc67daaf-3554-4c99-8c38-77aa1ac2464f	0.421225	0.454276	0.523659	0.619211		
20	P1000041.JPG	m1	79aa95f5-bf50-4962-8547-5ddde92129c3	0.886931	0.52488	0.959411	0.640909		
21	P1000042.JPG	ll	0987465a-f890-4212-9d5f-25b945fd173e	0.376303	0.437507	0.483726	0.608225		
22	P1000043.JPG	m1	01ce17d8-bcda-4425-8e64-0f06552e65ef	0.370657	0.394095	0.423835	0.479164		

，一共有 7 个字段，分别为 picName（图片名称），personName(识别出来的人的名字)，faceId（程序为每一个脸生成的唯一标识符），left（人脸框的左侧边框坐标），top（人脸框的顶部边框坐标），right（人脸框的右侧边框坐标），bottom（人脸框的底部边框坐标）。

建议：建议每次处理一个文件夹，即在步骤 2 中每次只导入一个文件夹，并且导入之后将之前的结果删除，将之前处理的文件夹“从 picasa 中删除”，保证每次处理的图片只来自于同一个文件夹。