مشخصه رفتاری تقسیم بندی تصویر جمعیت

1. معرفی

پلت فرم های برون سپاری افراد و شرکت ها را قادر می سازد تا به سرعت جمع آوری مقادیر زیادی از ورودی انسانی را انجام دهند. چالش این است که چگونه به نتایج حاصل از افراد غیرفعال از طریق محتوا اعتماد کنیم. کار ما سهمی از تحقیق روزافزون در مورد چگونگی موفقیت گروه های بزرگ جمعیت افراد غیر رسمی برای جمع آوری نتایج تفسیر تصاویر با کیفیت بالا است. در سراسر زمینه های مختلف، روش هایی برای کنترل کیفیت نتایج و به حداقل رساندن خطاهای انسانی پیشنهاد شده است. در میان آنها می توان دو دسته اصلی را مشاهده کرد. اولین دسته روش هایی برای طراحی وظیفه crowdsourcing مناسب است. به شرایط مناسب کاربران بالقوه، با استفاده از سرنخ های تصویری مناسب و دستورالعمل های جامع برای جلوگیری از ابهام، و نشان دادن مثال های خوب و بد نتیجه مورد نظر نیاز است، درخواست کنندگان می توانند شانس خود را برای به دست آوردن پاسخ های با کیفیت بالا از جمعیت افزایش دهند. دسته دوم روش های کنترل کیفیت می تواند در زمان اجرا مورد استفاده قرار گیرد. تکنیک های مورد مطالعه در این دسته عبارتند از: بررسی تخصصی دامنه و استفاده از انفصال، که در آن نتایج چند جمعیت به یک نتیجه نهایی می پیوندد. ناحیه کمتر تحقیق شده چگونگی استفاده از دانش در مورد رفتار کاربر برای ارزیابی نتیجه نهایی است.

اطلاعات جمع آوری شده برای حل وظایف دیداری رایانه ای مورد استفاده قرار گرفته است. تمرکز کار ما این است که مرزهای اشیاء در تصاویر (تقسیم بندی)، که یک گام بحرانی است برای بسیاری از وظایف دیداری کامپیوتر، از جمله جمع آوری اطلاعات در مناطق موردنظر، تطبیق اشیاء در تصاویر مختلف (ثبت نام)، دنبال کردن اشیاء (ردیابی)، و تمایز بین انواع مختلف اشیاء (طبقه بندی) را مشخص کند.

محتوای تصویر می تواند رفتار کاربر و کیفیت حاشیه نویسی های به دست آمده را تحت تاثیر قرار دهد - کاربر ممکن است زمان بیشتری برای رسم کردن مرزهای شیئ پیچیده، مانند مرز یک گل در مقابل جعبه نیاز داشته باشد. ممکن است از این دیدگاه استفاده کنید تا هزینه های جمع آوری رسم مرزهای شیء را کاهش دهید. یکی دیگر از ویژگی های آموزنده رفتار کاربر، تعداد کلیک هایی است که کاربر برای محدود کردن مرز انجام می دهد. ایده این است که بیشتر شکل های پیچیده نیاز به نقاط بیشتر برای گرفتن این شکل در جزئیات دارد. پیش ببینی شده است که تعداد کلیک های کاربر مربوط به دقت باشد که ممکن است از کاربران جمعیت هنگام درخواست آنها برای رسم مرزهای اشیا انتظار داشته باشید. ابزار طراحی همچنین می تواند موثر باشد.

ما یک مطالعه را برای بررسی رفتار طراحی کارگران پرقدرت در هنگام آشنایی و محتوای تصویر نا آشنا انجام دادیم ، که در تصاویر روزمره و بیومدیکال نشان داده شده است. ما اتصال را بین تکمیل کار دقیق و زمان اتمام کار تحلیل کردیم. همچنین تعداد کلیک های انجام شده کاربران برای رسم مرز با دقتی که این مرز کشیده شده بود را مقایسه کردیم. و آن را برای گسترش تجزیه و تحلیل رفتار جمعیت با ارائه اطلاعات ناآشنا بسیار ارزشمند یافتیم. مشکلاتی را که تا به حال به نظر می رسید نادیده گرفته شده اند،آشکار کردیم و ممکن است نیاز به بازنگری یا گسترش کلی فرضیه های مربوط به رفتار جمعیت داشته باشد. به طور خاص، ما تفاوت های شگفت انگیز در رفتار کاربر بر اساس آشنایی از محتوای تصویر را آشکار کردیم. نتایج ما اطلاعاتی را در مورد چگونگی جمع آوری اطلاعات که ممکن است برای وظیفه تقسیم بندی تصویر استفاده شود ارائه می دهد.

2. روش شناسی تجربی

آزمایش crowdsourcing تمرکز بر وظیفه تقسیم بندی تصویر برای جمع آوری طرحهای crowdsourced از مرزهای اشیاء دارد، ما از ابزار Annotation آنلاین دسترسی آزاد LabelMe استفاده کردیم. آن را به Amazon Mechanical Turk متصل کردیم، یک بازار اینترنتی که ما کاربران را استخدام کردیم که قبلا حداقل 100 کار (HIT) انجام داده بودند و حداقل یک امتیاز تایید 92٪ را نگهداری می کردند. ما تمام HIT های ارسالی را پذیرفتیم. کاربران را با دستورالعمل های جزئی در مراحل کارهای طراحی و تصاویر نشان داده شده آنها از حاشیه نویسی خوب و بد ارائه دادیم. زمان تکمیل کار مکانیزم پایان را به دست آوردیم.

برای ارزیابی کیفیت نقاشی، از نسبت همپوشانی منطقه، یک معیار ارزیابی استاندارد برای اندازه گیری دقت تطابق تقسیم بندی با مرز شیء "واقعی" (استاندارد طلا) استفاده کردیم. نسبت همپوشانی شمار پیکسل های معمولی را برای هر دو ناحیه توضیح داده شده واقعی که در ترکیبی از مناطق هستند شمارش می کند |A∩B/A∪B|، (A نشان دهنده مجموعه ای از پیکسل ها در منطقه واقعی است و B نشان دهنده مجموعه ای از پیکسل ها در منطقه یادداشت شده است). رنج نمرات از 0 به 1 با نمره بالاتر بازتاب بیشتر شباهت و دقت بهتر دارد.

ما از دو کتابخانه تصویری معروف رایگان که شامل دو نوع تصاویر هستند استفاده می کنیم: " تصاویر روزانه "از اشیاء آشنا مانند پرندگان و ساختمانها ، و تصاویر زیست پزشکی از محتوای نسبتا نا آشنا، مانند سلول های عضلانی و مقاطع عرضی آئورت قلبی (به ارقام نگاه کنید) علاوه بر داده های تصویری خام، کتابخانه ها شامل، هر تصویر نسخه های چندگانه از مرزهای شیء رسم شده توسط کارشناسان (دانشمندان در مورد تصاویر زیست پزشکی) است. برای به دست آوردن یک استاندارد طلایی تک برای مرزهای شیء "واقعی"، ما این مرزهای تخصصی رسم شده را با استفاده از رای اکثریت در سطح پیکسل (اختصاص یک پیکسل برای تعلق به این شیء زمانی که حداقل نیمی از حاشیه نویسی متخصص پیکسل را به جسم اختصاص می دهد) به هم متصل کردیم.

برای هر کدام از 405 تصویر در مجموعه داده های ما، پنج طرح crowdsourced جمع آوری شده است. هر کاربر مرز شیئ اصلی تصویر را با قرار دادن بسیاری از نقاط کنترل به عنوان she یا he که لازم به نظر می رسد ردیابی می کند. رابط کاربری پیوسته نقاط کنترل را با خطوط مستقیم متصل می کند. این کاربر با کلیک کردن بر روی نقطه شروع، ردیابی مرز را به پایان می رساند. ما به تعداد نقاط کنترل به عنوان "تعداد کلیک های کاربر" اشاره می کنیم. از رای اکثریت در سطح پیکسل استفاده کردیم تا پنج جمعیت جمع شده را به یک طرح کلی جمع آوری کنیم (به عنوان مثال، اختصاص یک پیکسل متعلق به جسم فقط اگر حداقل نیمی از کاربران آن را به جسم اختصاص دادند). ما دقت طرح کلی محاسبه شده را با اندازه گیری توافق آن با طرح استاندارد+ طلایی -expert established با استفاده از معیار همپوشانی متریک تجزیه و تحلیل کردیم.

3. نتایج

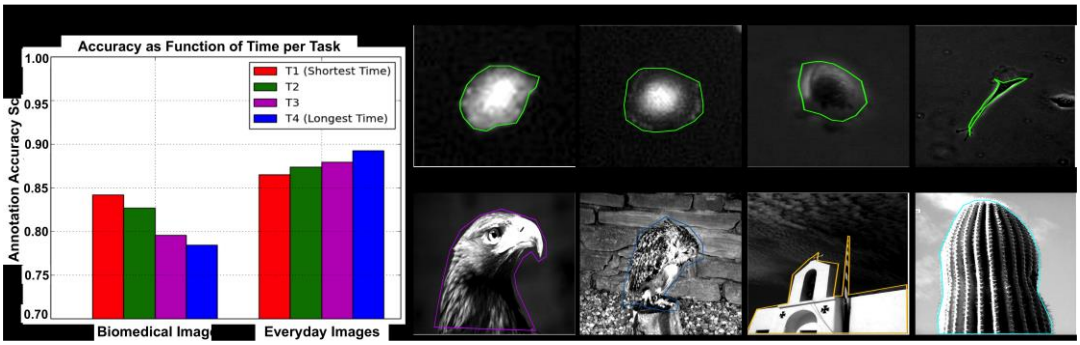
در آزمایش ما، 90 کارمند منحصر به فرد 2025 نقاشی را به خود اختصاص دادند. ما در اینجا رفتار کاربر جمعیت را با رابطه بین کیفیت یادداشت و تعداد کلیک های کاربر رسم جمعیت بخوبی زمان لازم برای حاشیه نویسی

مشخص می کنیم. میانگین زمانی که کاربران برای رسم یک طرح کلی شیء در نظر گرفتند 46 ثانیه بود و میانگین آن 30 ثانیه بود (25 درصد 21ثانیه و 75 درصد 56ثانیه). متوسط زمانی که کاربران برای رسم یک طرح کلی در یک تصویر بیومتریک به کار می بردند، 26 ثانیه بود (25 درصد 18 ثانیه و 75 درصد 46 ثانیه) و در تصویر روزانه 32 ثانیه (25 درصد 48ثانیه و75درصد 80 ثانیه).

تعداد متوسط کلیک کاربر 27 بود (25 درصد 20تا و 75 درصد 38تا)، 27 مورد برای تصاویر پزشکی (25 درصد 21 و 75 درصد 37 تا کلیک کاربر) و 29 برای تصاویر روزانه (25 درصد بارز 17 و 75 درصد 43 تا کلیک کاربر).

3.1 دقت به عنوان عملکرد زمان تکمیل

برای بررسی رابطه بین کاربران زمانی که به تصاویر یادداشت پرداختند و کیفیت کار آنها ما داده های زمان تکمیل را به چهار دسته تقسیم کردیم، T1 برای 0-25 درصد، T2 برای 25-50، T3 برای 50-75، و T4 برای 75-100. سپس دقت نتایج حاشیه نویسی برای تصاویر بیومدیکال و روزانه را با توجه به این گروه های تقسیم شده تجزیه و تحلیل کردیم (شکل a 1). هنگام مقایسه دقت حاشیه نویسی تصاویر بیومدیکال در مقابل تصاویر روزمره، ما یک الگوی مخالف را از تغییر در نمرات دقت با افزایش زمان اتمام مشاهده کردیم: به عنوان مثال زمان تکمیل، کار افزایش یافته کیفیت حاشیه نویسی تصویر در تصاویر بیومدیکال را بدتر کرده است. با این حال، در تصاویر روزانه، یک کاربر زمان بیشتری صرف کرد، نتایج حاشیه نویسی تصویر دقیق تر به دست آمد.



شکل 1 (الف) دقت علامت گذاری به عنوان تابع زمان تکمیل کار. نمرات به چهار دسته تقسیم می شوند (T1-T4) و برای تصاویر بیومدیکال و روزانه به طور جداگانه گزارش شده است. (ب) حاشیه نویسی نشان می دهد که نتیجه رسم جمعیت ازهر یک از T1 T2، T3، و T4 برای تصاویر زیست پزشکی و روزمره. برای هر شی، 17 کلیک کاربر استفاده شد.

3.2 دقت به عنوان یک عملکرد از کلیک های کاربر

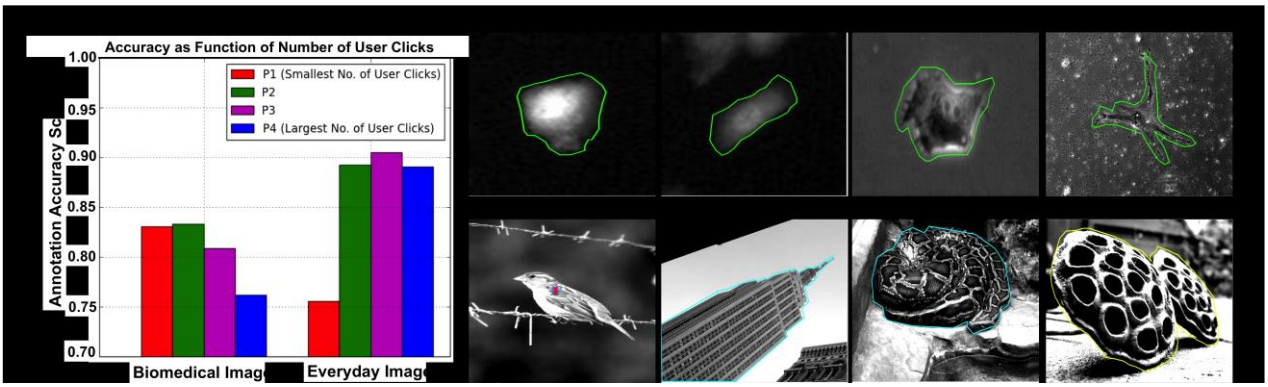
ما رابطه بین دقت کاربر و تعداد کلیک های کاربر را مورد بررسی قرار دادیم (شکل 2 a).همانطور که درتجزیه و تحلیل قبلی، نتایج حاشیه نویسی را به چهار دسته تقسیم کردیم، P1 برای اولین تقسیم،P4 برای آخرین. همچنین هنگام مقایسه دقت علامت تصاویر زیست پزشکی و روزانه ویژگی های متفاوتی پیدا کردیم. برای داده های زیست پزشکی، کمترین نمرات دقت برای تقسیم بندی هایی که بیشترین تعداد کلیک های کاربر را دارند (P4) به دست آمده است. برای تصاویر روزانه، با این حال، بیشترکلیک کاربر به معنای دقت بالاتر است (P2-P4). تقسیم بندی تصاویر روزانه به دست آمده با کمترین کلیک کاربر (P1) در داشتن کمترین نمرات دقت و در (یا وزن کمتر)یک روش رأی گیری که حاوی حاشیه نهایی است میتواند مورد توجه قرار گیرد.

4. بحث و نتیجه گیری

ما شگفت زده شدیم که زمان بیشتری منجر به نتایج دقیق تر برای تفسیر محتوای آشنا در تصاویر روزانه و نتایجی با دقت کمتر برای محتوای نا آشنا از تصاویر biomedical شد. ما تصاویر بصری را بررسی کردیم تا نتیجه بگیریم (شکل 1 b). فرض می کنیم که کاربران که زمان بیشتری برای انجام دقیق هر نقطه کنترل صرف کرده اند در امتداد مرز شی وجود داشت. این مرز برای اشیاء روزمره قابل درک است، اما برای تصاویر بیومدیکال قابل درک نیست و مبهم است(شکل 2 b). همچنین شگفت زده شدیم که کلیک های بیشتر کاربر، یعنی نقاط کنترل مرزی، منجر به روندهای مختلف مشاهده شده از کیفیت یادداشت با محتوای تصویر آشنا و نا آشنا شد. بازرسی دیداری از نتایج نشان می دهد که مرزهای اشیای زیست پزشکی بسیار پیچیده است افزایش تعداد نقاط مرزی به طور کلی در بهبود دقت بخش بندی ناکافی بود (شکل b2).

کار ما، تعیین اینکه چه چیزی از کاربران جمعیت در زمان و مکان انتظار می رود، تعداد کلیک برای دو مجموعه تصویر چالش انگیز که محتوای تصویر آشنا و آشنا را نشان می دهند تسهیل می شود. امیدواریم که این درک افراد را تشویق کند تا روش های کنترل کیفیت عمومی تر یا سیستم های طراحی مبتنی بر وب را برای سیستم های ذخیره سازی خود تشویق کنند. علاوه بر این، ما این را امیدواریم که همکاری های بین رشته ای، آینده را با

برجسته کردن نحوه تجزیه و تحلیل در مورد تصاویر ناآشنا تشویق خواهد کرد (بیومدیکال) و راه هایی برای بهبود درک ما در مورد رفتار کلی جمعیت را نشان می دهد.



شکل 2 (a) دقت علامت گذاری به عنوان تابع تعداد کلیک های کاربر در هر کار. نمره ها به چهار دسته تقسیم می شوند (P1-P4)و برای تصاویر بیومدیکال و روزانه به طور جداگانه گزارش شده است. (ب) حاشیه نویسی نماینده نشان دهنده یک نتیجه رسم جمعیت از هر یک از P1، P2، P3 و P4 برای تصاویر زیست پزشکی و روزمره است. زمان انطباق برای هر تصویر از 42 تا 56 ثانیه متغیر است.

REFERENCES

M. Allahbakhsh, B. Benatallah, H. R. Motahari-Nezhad, A. Ignjatovic, and E. Bertino. 2013. Quality Control in Crowdsourcing Systems. *Internet Computing* 17, 2 (2013), 76–81.

S. Alpert, M. Galun, R. Basri, and A. Brandt. 2007. Image Segmentation by Probabilistic Bottom-Up Aggregation and Cue Integration. In *Proceedings of the Conferece on Computer Vision and Pattern Recognition (CVPR), 8 pages*.

L. Biewald and C. Van Pelt. 2010. Distributing a task to multiple workers over a network for completion while providing quality control. US Patent App. 12/817,946. (June 2010).

A. Carlier, V. Charvillat, A. Salvador, X. Giro i Nieto, and O. Marques. 2014. Click'n'Cut: Crowdsourced Interactive Segmentation with Object Candidates. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*. ACM, 53–56.

D. Gurari, D. Theriault, M. Sameki, and M. Betke. 2014. How to use level set methods to accurately find boundaries of cells in biomedical images? Evaluation of six methods paired with automated and crowdsourced initial contours. In *Interactive Medical Image Computation Workshop (IMIC)*.

P. G. Ipeirotis, F. Provost, and J. Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 64–67.

A. Kulkarni, M. Can, and B. Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 1003–1012.

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1–3 (2008), 157–173.

J. M. Rzeszotarski and A. Kittur. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 13–22.

H. A. Sampath, R. Rajeshuni, and B. Indurkhya. 2014. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. 3665–3674.

A. Sorokin and D. Forsyth. 2008. Utility data annotation with Amazon Mechanical Turk. *Urbana* 51, 61 (2008), 820.

S. Vijayanarasimhan and K. Grauman. 2009. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *Conference on Computer Vision and Pattern Recognition*. 2262–2269.