Transformative Crowdsourcing: Harnessing the Power of Crowdsourced Imagery to Simplify and Accelerate Investigations

Minwoo Park, TaeEun Choe, Andrew Scanlon, Allison Beach

Research and Development Services, ObjectVideo
{mpark,tchoe,ascanlon,abeach}@objectvideo.com

Abstract. When an emergency or terrorist attack such as Boston Marathon bombings occurs law enforcement and the intelligence communities are often inundated with large quantities of visual media. This media comes from captured sources such as cell phones and computers, from surveillance and security cameras, as well as from public social sources such as Flickr or YouTube. Once collected, analysts must comb through terra-bytes of image and video data looking for evidentiary leads. Automated computer vision algorithms help to pare down the search space; however analysts still spend large amounts of time manually reviewing hundreds of hours of video and thousands upon thousands of images searching for the needle-in-a-haystack lead. To ameliorate these challenges, we propose "Transformative Crowdsourcing", a crowdsourcing system for sorting, managing, and reviewing large quantities of visual media. Investigators may use this application to seek help resolving a case through the either public or private (internal) crowdsourcing of leads, image/video searches, or for seeking more information about a specific time, place, or person.

1 Introduction

While the idea of an army of multi-talented resources sounds like the perfect workforce, crowd sourcing does have its challenges. First, the quality of the results cannot be guaranteed. Some annotators are excellent resources, others may be less diligent or capable, and some may purposefully attempt to corrupt the results. Applying adequate quality control is a significant challenge to using crowd-generated results. Secondly, crowds are typically composed of people with varying skills and expertise. Tasks must be presented in an intuitive manner to elicit the best results from users. Finally, enticing people to participate is another crowdsourcing challenge. The interface must be engaging to motivate users to participate, particularly if they are not to be financially compensated.

Transformative Crowd sourcing turns terra-bytes of raw visual media into searchable metadata in a fraction of the time of traditional manual analysis. Analysts may quickly and more accurately perform tasks such as geo-locating media, identifying/following a person-of-interest through time and space, and creating links between people and entities simply by leveraging the power of the crowd.

Within computer vision, crowdsourcing has proven especially useful in large-scale image label collection, as many computer vision algorithms require substantial amounts of training data. However, there are technical challenges: 1) *motivating workers to complete tasks*, 2) *designing a task*, 3)*ensuring high quality results*, and 4) *deploying the task efficiently*.

We propose a system that addresses above challenges for a set of predefined tasks but not limited to: 1) Find a subset of imagery that contains a suspect. 2) Find space and time trajectory of a suspect. 3) Find a geo-location of a given multimedia data.

2 Transformative Crowdsourcing

We describe our proposed system in a context of finding a suspect. However, our proposed system can adapt to the different tasks. Note that the robust detection of a suspect in a set of imagery enables algorithms to build deep association graphs between people, objects, and activities in the visual media, which can then drive the semi-automated creation of a suspects spatial and temporal storyboard.

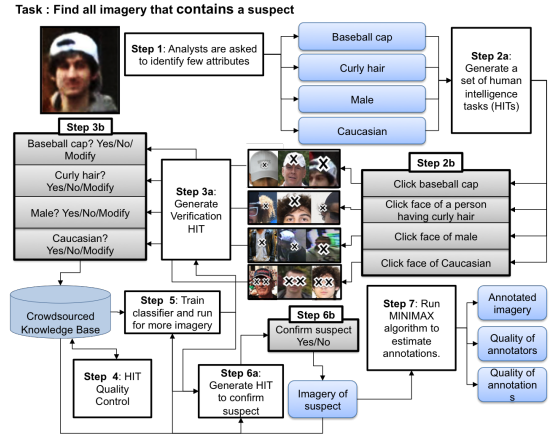


Fig. 1: *A block diagram of our proposed interactive crowdsourcing system.*

The Fig. 1 shows a block diagram of the proposed system when the task is Finding a suspect. In step 1, analysts (experts) are asked to annotate a set of attributes about a suspect in a set of imagery. E.g., white baseball cap, curly hair, gray colored hoodie, and Caucasian male.

In step 2a, our proposed system generates the first set of human intelligence tasks (HIT)s. The type of HIT produced here is a pointing HIT where the annotators are asked to click. With a goal of finding suspect in imagery, but not knowing what they are doing, annotators are asked to click baseball cap on a person, face of person with a curly hair, face of a male, and face of a Caucasian, independently.

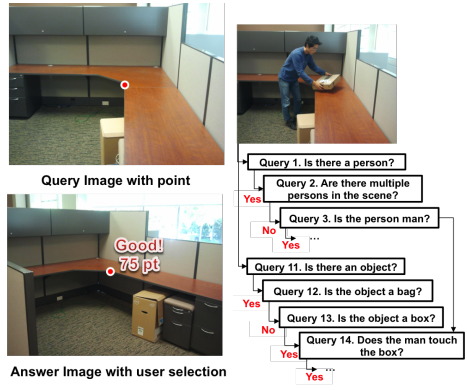
In step 2b, annotators perform such HITs. For a general goal, this pointing HIT can be used to make point correspondences between two images. First, a set query point is automatically generated from feature detection algorithms such as SIFT, SURF, ORB, etc. The annotators are asked to click the corresponding point in a different image taken by other camera. The matched points can be used for calibration and geo-registration of sensors. Fig. 2a shows such a HIT.

In step 3a, verification HITs are produced to verify the results of the step 2b. The step 3a is invoked whenever the results are available from the step 2b.

Therefore it is possible for the steps 2b and 3b co-exist. The HIT in the step 2b corresponds to "creation task" in a human computing process and the HIT in the step 3b corresponds to "decision task in a human computing process [1]. The outcome of the steps 2b and 3b can boost the quality of annotations using human intelligence. Specifically, with the goal of finding a suspect, annotators are now asked to answer if they see a white baseball cap in an extracted sub-image centered at the previously clicked locations by other annotators in the step 2b. This process obviates the need to scan entire portion of an image or a video since the region of interests are already identified by the step 2b.

For a general goal, this HIT can be augmented by the Restricted Turing Test (RTT) query engine automatically. In the beginning, the questions are simple asking if the scene contains a person, car, building, or object. However queries will be changed based on the users answers. For example, when the user answered, "Yes, there is a car." Then HIT will ask deeper questions such as, "What is the color of the car?" and "Can you find the car in this image?" Fig. 2b shows one example of such RTT in our system.

In step 4, the quality of the annotations for all annotators are measured and controlled. If the annotations made in the step 2b are rejected in the step 3b by other annotators, annotators who generated rejected annotations will be ranked lower. If the annotations made in the step 2b are accepted in the step 3b by other annotators, annotators who generated accepted annotations will be ranked higher. By the human intelligence, the performance of annotators can be measured in this stage. As more data are generated, pseudo ground truth can be identified by minimax algorithm [2] and the annotators can receive immediate positive or negative feedback on their annotations (Figure 2a). The minimax algorithm is a variant of Dawid-Skene estimator [3] algorithm that enables 1) estimation of annotators' ability, 2) HIT difficulties, 3) balancing annotators' ability and the HIT difficulty, and 4) inferencing of ground truth.



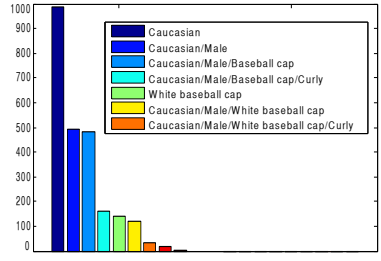
(a) Pointing HIT (b) Yes/No HIT

Fig. 2: *(a) Point HIT where annotators are asked to click corresponding point (red dot) of the query image in the answer image. (b) Yes/No HIT (RTT) where annotators are asked to answer simple questions and the subsequent questions are displayed depending on the previous answers.*

In step 5, we train a classifier for each attribute to expedite the annotation process. The results of the classification are further verified in the step 3b by the human intelligence. Then this process is repeat until convergence. This is an iterative crowd-enabled active learning process similar to [4] for building high-precision visual classifiers. The verified results in the step 3b seed a classifier,

which is then iteratively trained by active querying of the annotators. These annotators actively refine the classifiers at every iteration by answering simple binary questions about the classifiers detections. The advantage of this approach is to expedite and aid the annotations procedure in step 2b. This obviates the need to perform the step 2b for an entire multimedia set to provide input for the step 3a.

In step 6a, a new HIT is generated to confirm if the imagery contains the suspect. This process obviates the need to scan entire portion of an image or a video since the region of interest is already identified by the previous steps which maximize ergonomics of the task. The set of imagery used for this HIT can be collected by using a set of conditions made out of the attributes. E.g., the set can be retrieved by collecting all imagery that contains the white baseball cap, curly hair, and Caucasian male. The results of step 6b are fed back to step 5 to train a classifier to detect the suspect then the results are further verified in the step 6b. This process is repeat until convergence. This expedites the process to provide input for the step 6a. Finally, in step 7, we perform the minimax algorithm [2] again to refine all of the annotations.



(a)

(b)

Fig. 3: *(a) Number of instances per a set of tags. (b) Some of the corresponding results.*

3 Results and Conclusion

Figure 3 shows proof of concept results from our system. Note that a set of simple attributes is enough to pin-point the suspect. The entire pipeline is still in development and the results provided here are small part of the entire pipeline. We believe our proposed system can address the aforementioned challenges discussed in Section 1 for predefined set of tasks.

References

1. Little, G., Chilton, L.B., Goldman, M., Miller, R.C.: Turkit: Human computation algorithms on mechanical turk. In: Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology. UIST '10, New York, NY, USA, ACM (2010) 57–66
2. Zhou, D., Liu, Q., Platt, J.C., Meek, C.: Aggregating ordinal labels from crowds by minimax conditional entropy. In: Proceedings of the 31st International Conference on Machine Learning (ICML). (2014)
3. Dawid, A.P., Skene, A.M.: Maximum likeihood estimation of observer error-rates using the em algorithm. (Applied Statistics) 20–28
4. Patterson, G., Horn, G.V., Belongie, S., Perona, P., Hays, J.: Bootstrapping fine-grained classifiers: Active learning with a crowd in the loop. In: NIPS Workshop on Crowdsourcing: Theory, Algorithms and Applications, Lake Tahoe (2013)