Kevin Luo
Prof Ethan Whittet
Advanced Writing in the Tech Professions
3/12/2016

Developments in Facial Landmarking

The Problem of Facial Landmarking

In computer vision, facial landmarking is the problem of identifying the components of a face given a photograph of the it and a series of coordinates that define its bounds. Generally, this involves producing a list of coordinates on the face that correspond to outlines of prominent facial features, eg. eyes, nose, and mouth. The problem has been tackled by many scientist and engineers in the field of computer vision, and has seen much progress and development recent years.

Solving the problem quickly and accurately has extensive applications in industry. Identified landmarks on a face can be used to recognize individuals across different photographs. Social networks and security software both rely heavily on this ability to track and monitor individuals. Similarly, much work is being done in the the ability to recognize facial expressions, and many applications of this are also dependent on landmarked faces. In a recent press release, Disney Research released a software package called FaceDirector, which allows an editor to seamlessly synthesize and apply an actor's facial performance to an existing scene based on multiple takes.

The technology is now being employed in end-user applications, particularly on smartphones. Snapchat's most recent update allows users to apply one of many filters that react to the location of and expression on a user's face. L'Oreal's "Makeup Genius" digital makeup application software, which uses the identified landmarks to apply makeup to a user's photo.

The state of the art algorithm for landmarking photos of faces is currently an application of machine learning, requiring a collection of images of faces is provided with a list of the corresponding landmarks according to a consistent scheme, usually provided by a human. This data and the set of photos are known as the ground truth. The photos and accompanying annotations are processed by an algorithm to produce a model that can be used to guess at the landmarks of any new photograph, and the efficacy of the model that results is the product of both the algorithm that generated it, and the quality of the ground truth data.

As the field of facial landmarking has developed, it has come to rely more and more specific tools, forgoing algorithms that work in more general cases to favor those more specific to the challenge of facial landmarking.

The Active Appearance Model

An early implementation of algorithms that solved this problem was described by Tim Coote's work in his work on Active Appearance Models, which was the standard of face detection until the mid 2000's. It also provided a basis for much of the terminology that is now commonly used in the practice of facial landmarking. Coote's technique involved using gradient vectors generated from an image as coordinate system from which landmarks could be placed based on template matching algorithms. His technique was shown to work well with a few landmark points, up to 8. Many practical applications however, required many more, and many more landmark points made the algorithm unfeasible.

Coote's work was a vast improvement, and possibly the defining factor in making facial landmarking feasible. Earlier attempts to define this problem used a data structure called active contour models, or snakes colloquially. These models were capable of matching simple contours based on adjacent template matching. They were ideal for isolating discrete sections of an image, but were less useful when applied to nonlinear tasks like identifying the landmarks on a face. They were an attempt to borrow a general algorithm from machine vision to attempt this task.

Cascading Regression Functions

Paul Viola and Michael jones introduced a seminal technique in their 2001 paper, which became the state of the art landmarking for years to come. Their crux of their novel technique was in the use of a series of cascading regressors, each adjusting the result of the landmarking estimation in a specific way, and each depending on the output of the regressor from higher up in the chain. This chain system drastically reduced the complexity of each regressor, allowing many more feature vectors to be used, attaining a high degree of accuracy and speed.

As of the computer vision and pattern recognition (CVPR) 2014 expo, the best implemented algorithm using this cascading regressor system was described in a paper by Kazemi, Sullivan et Al and heralded as the "One Millisecond" solution. The most common algorithm used today using a cascading regression model, in which an average guess for the landmarking is generated as the average of the training set, centered on a provided bounding box. From there, the a function is used to progressively refine the model based on parameters generated from the training set.

The solution presented in this paper expands on the ideas of Active Appearance Models by using a cascading function instead of a regression function. Previous work used features extracted during the learning phase are used in a regression function on each image
This method is likely to produce an optimal solution, but at the heavy cost of time and memory, especially for models with many landmarks

Instead, the proposed technique uses a cascading algorithm, iteratively adjusting the model based on the extracted features until a optima is reached. The algorithm creates a series of regressors each minimizing the difference in squares, each regressor limiting itself in the features it modifies. In this way, each regressor is limited in size, improving overall performance.

Cascading Regressor Function Limitations

Two issues immediately emerge from the this model of facial landmarking. The first revolves the high degree of initialization dependency inherent in using a cascading. Depending on the model based on the average of the training set, the algorithm may settle on a landmarking configuration what is locally optimal, but not the globally optimal solution.

Some algorithms use a element of randomization to combat this, either running the algorithm using a variety of different initializations, or by making nonoptimal feature transforms on occasion. Others employ an element of randomness that relies on a computed metric of accuracy during each iteration, making progressively less erratic guesses as the estimate converges. Of course, these are all strategies that are commonly used to combat the major weakness of greedy search algorithms, their tendency to converge on local extrema. As in general programming, these techniques are effective only to a certain point. A degree of initialization independence is necessary avoid the pitfalls of local extrema

The second issue is the limitations regarding posture. Only variations seen in the training set will produce parameters that can be adjusted by the algorithm as it searches for a landmarking solution. The resulting solution has limitations in the robustness; it frequently has difficulties with recognizing variations in posture that are not seen in the training set. Other issues, such as partial face occlusion also heavily hinder the effectiveness of the algorithm, and adjustments for this are difficult to make in both the algorithm and the data set.

Coarse to Fine Shape Searching

In the most recent CVPR expo, Zhu et Al provided a novel approach to the problem of landmarking that despite its complexity and specificity, has been shown to consistently outperform the best algorithms to date. Zhu's paper could possible represent an innovation as seminal to the field as Kazemi's was fifteen years ago.

Zhu's work also improves on speed, since the early stages of the algorithm can use more efficient types of algorithms designed to detect vague shapes quickly and accurately, while using those models to then obtain more accurate pictures of detailed landmarks, such as identifying general eye regions. When the need more detailed landmarking arises in the later stages of the algorithm, more detailed, but computationally intensive algorithms can be used.

The algorithm uses a number of simplified models from the training set as candidate shapes for initialization. Although initialization is random in Zhu's algorithm as well, it takes place on a more abstract level, matching coarse shapes and giving landmark estimates distributions instead of fixed locations. Iteration takes the form of feature transformations based on generalizations of the faces seen in the training set. A number of candidate transformations are generated and considered for each iteration.

To address the issue of robustness of posture recognition, Zhu et al proposed that candidate shapes and transformations would be based on some of the postures seen in the dataset. Since these postures are not averaged, they retain their individual combination of features, Zhu et al also note that Kazemi's cascading regression algorithm includes a shape indexing step, which they claim is able to improve the speed of their algorithm if implemented.

Performance

Zhu et al performed standard tests on their algorithm, using the standard IBUG dataset of over 10,000 annotated face images. Since their method is open with regards to a feature detection algorithm, they employed a combination that was deemed to be the most practical in the their test sessions, using Binary Robust Independent Elementary Features (BRIEF) for the initial iterations and in a few iterations and Scale-Invariant Feature Transform (SIFT) for the final shape search. BRIEF is a faster shape searching algorithm and SIFT, which is more accurate. With both this configuration, and another made with exclusive use of SIFT, deemed the most optimal, their results compared favorably to the ERT results.

Zhu et al compared the results of their course to fine shape searching using a cumulative error comparison. Specifically, they took a specific measurement across a large dataset specifically for testing, measuring the difference between the located pupils of the landmarks generated by the trained model and of the human generated ground truths.

On the hardware that Zhu et al used to test the algorithms, coarse to fine shape searching was found to be both faster and more accurate that the standard ensemble regression trees. The team found noticeable increase in speed. Using two comparable MATLAB implementation and an intel i5 processor, the team observed 40ms per frame. This is comparable to the 28 ms per frame on a cascaded regression tree per initialization. In most implementations however, a number of random initializations is tried, usually around 5.

Their algorithm shows evidence of increased accuracy as well. The average inter pupil error across the entire IBUG dataset was found to be 5.99%, compared to 6.31% for the most successful ensemble regression tree implementation.

Further inquiry

For the most part, current facial landmarks are part of a chain of tools that start with facial identification, producing a bounding box that finds a face for the landmarker to identify. Zhu et al's work indicates a trend towards a more stratified technique using different algorithms to identify general facial characteristics before identifying specific landmarks. As this field matures, it is likely that even more specified algorithms will be introduced at different levels of the landmarking process to find the optimal procedure for landmarking.

In a more recent development Zhang et al have proposed an innovation in landmarking that uses neural networks to account for factors that aid in identifying landmarks. Using existing algorithms, they identify characteristics of a face, such as gender, age, presence of glasses, and any occlusions to the face to improve their estimation of landmarking. Their aim is to make a more robust landmark detection algorithm through the use of regressions that are dependent of these factors.

The general trend is towards techniques that are more and more specialized to the specific challenge of facial landmarking. Engineers have been moving away from general contour and shape recognizers towards specific made tools to solve this difficult problem, and their results have been steadily improving as a result.

Works Cited

"Incredible Software from Disney Research Seamlessly Blends Faces from Different Takes." *No Film School*. N.p., n.d. Web. 14 Mar. 2016.

Zhang, Zhanpeng, Ping Luo, Chen Change Loy, and Xiaoou Tang. "Facial Landmark Detection by Deep Multi-task Learning." *Computer Vision – ECCV 2014 Lecture Notes in Computer Science* (2014): 94-108. Web.

Milborrow, Stephen, and Fred Nicolls. "Locating Facial Features with an Extended Active Shape Model." *Lecture Notes in Computer Science Computer Vision – ECCV 2008* (n.d.): 504-13. Web.

Kazemi, Vahid, and Josephine Sullivan. "One Millisecond Face Alignment with an Ensemble of Regression Trees." *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014): n. pag. Web.

Zhu, Shizhan, Cheng Li, Chen Change Loy, and Xiaoou Tang. "Face Alignment by Coarse-to-fine Shape Searching." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): n. pag. Web.

Viola, Paul, and Michael Jones. "Rapid Object Detection Using a Boosted Cascade of Simple Features." *IEEE Xplore*. N.p., n.d. Web. 14 Mar. 2016.